Genetic Architecture underlying rosette morphology quantified by Computer Vision

Genotype to phenotype mapping on three *Arabidopsis thaliana* L.(Heyn) experimental populations

Odín Manuel Morón García

A thesis presented for the degree of

Doctor of Philosophy

Institute of Biological, Enviromental and Rural Sciences

Aberystwyth University

Wales - United Kingdom

05/03/2018

Genetic Architecture underlying rosette morphology quantified by Computer Vision

Genotype to phenotype mapping on three *Arabidopsis thaliana* L.(Heyn) experimental populations

Odín Manuel Morón García

Abstract

Arabidopsis thaliana(L.)Heyn is a mostly Eurasian species of Brassicaceae with a rosette habit during the vegetative phase. At preliminary experiments, it has been observed that variation in rosette morphology in the juvenile stage, from seedling to flowering, is ecotype and environment specific phenotype .

I have examined the genetic basis of rosette architecture putting together whole-rosette high-throughput phenotyping and "phenotype to genotype mapping". Each rosette has been measured for Shape Descriptors derived from Digital Geometry using Computer Vision. Shape Descriptors have been use as traits for Association Mapping (GWAS) and Linkage mapping in three experimental populations: Natural Accessions, Recombinant Inbred Lines derived of a Cape Verde Island x Argentat cross, and Recombinant Inbred Lines from a Multiparent Advanced Generation Intercross (MAGIC).

GWAS and Linkage mapping found four potential QTLs during an initial scan. From MAGIC fine-mapping population, 41 potential Quantitative Trait Loci were found associated with rosette global architecture. I hypothesized that genes that integrate developmental response to environment (*Erecta*, *PhyB*) have influenced the developmental canalization of rosette morphology in juvenile plants.

Acknowledgements

As in any important project in a person's life and career, the journey towards obtaining a Ph.D. degree is full of people that provide required support, help and affection. It is impossible to name all these people in my journey to this attainment, however I will try my best in the next lines, apologising if I have missed unintentionally any name out.

Firstly, I would like to express my gratitude to my supervisors, Anyela Camargo-Rodríguez and John Doonan, for their patience, tolerance and support, regardless of all the stress moments and tensions associated to a Ph.D. I would like to especially thank John Doonan for bearing me in mind and for inviting me over to Aberystwyth for a second time. Also, the group at the National Plant Phenomics Centre has been exceptionally lovely and caring over the years, in particular Candida Nibau for the many coffees and walks along with conversations, Roger Boyle for his advice in computer vision and evenings under the sun with a beer, and Kevin Williams for his advice and support. There were also many others from the NPPC which played a relevant role in the development of this thesis, like Fiona Corke, Gina Garzón, Despoina Dadarou, Andreu Herrera, Karen Pricey, Jason Brook, as well as many others such as the visitors to our department. Also, I would like to appreciate the support of fellow researchers, like Hannah Dee, Richard Webster, Amanda Clare, Christine Zarges, Adil Mughal, Martin Swain, and many others whose conversations and modules were very rewarding.

My years in Aberystwyth allowed me to share some terrific experiences with many friends and colleagues. I want to express my sincere gratitude to Morgane, Marta, Eve, Tom, Emil, Alberto, Maciej and very specially to Rakesh Bhatia and Olga García for the endless hours of talking over our fantastic dinners and hiking. The other half of my friends that made my "Aber" experience great was the "Plas Tudor" crew, including Sarah, Florian, Kat, Adhemar and the bunch of International Politics friends (Prithvi, Matt, Katherina, Marcello & Gabriela,...). The uncountable hours sharing experiences, discussions and food made me a better person.

Secondly, the route to the culmination of this Ph.D. is populated with people that contributed in a way or other in the development of my academic career. I am happy to mention in a "backward time mode" all those that come immediately to my mind but do apologise if I may have forgotten to mention someone. My experience working for PSI in Brno, Czech Republic, was a life-changing experience, on one side thanks to my supervisors Klara Panzarova and

Michal Ŝicner and colleagues like Sharmila Madhavan and Ondrej Panzar, but for my personal experience I would like to cite especially Gerardo and Susana for so many things that I have no space to mention.

From the Spanish side of my career, I need to give thanks to Bruno Contreras, Juan Antonio Aguilera, Hilario Ramírez and Juan Pedro Camacho, for their guide and knowledge transmission. From my time at Granada University onwards, the support of my friends, specially Javier Valverde, has been unforgettable. Names that cannot be left out of this acknowledgement are Sito Ariza, Angela Sánchez, Ana and Kina García, Rafa Picazo and Iraitz Montalban among many others. Similarly, I need to mention Javi, Jose and Juan Redondo, Miguel Romero and the rest of my group from Cartaya, Spain.

At an atemporal scale, I am very grateful to my closest circle. I am really thankful to María Velasco, Lucie Buřilova and Pilar Martínez for all the moments we have shared, the love, support and patience, including their families (Paco, Bea, Marcos, Sofia, Ŝarka, Marie & Miloŝ, and Lukas). You really made my life way better.

Finally, my family, my dad, Juan, my mum, Loli, my auntie, Concha, my brother, Aitor, and my grandparents (Manuela & Manuel, Catalina & Juan), have been the strongest support, always helpful, always encouraging, providing feedback to any of my decisions of any kind without hesitation and with all confidence. Thanks y'all!

Odín Morón García.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General Plant Shoot Architecture

The aerial parts of plants exhibit a myriad of architectures with different degrees of complexity (Reinhardt, 2002). They range from little rosettes to big trees, shrubs or bushes, from plants with a primary growth axis, e.g maize, to complex branching pattern, e.g ferns, from simple leaves to complex leaflets, e.g tomato, and a plethora of leaf positioning patterns and leaf shapes (Teichmann and Muhr, 2015).

In spite of this complexity, at a lower level, shoots are composed of repetitions of simple units, called phytomers, consisting of an internode, i.e a portion of stem, and a node with an axillary leaf and a bud (meristem). The tissue responsible for building the phytomers is the apical meristem, e.g Shoot Apical Meristem (SAM) or Axillary meristems in the axil of leaves. Meristems control the growth and development of new phytomers, organize the leaf initiation, and eventually the progression to reproductive structures. The phytomer growth pattern is called a *plastochron*. The size of phytomers determines not only whole plant size, but the general structure of the plant. For example, long phytomers result in tall plants with sparse leaves between phytomers, like maize, and those with short internodes result in rosettes, e.g Arabidopsis or Tobacco. Meristem also controls the position and angle of leaves, resulting in opposite or alternate leaves pattern, and the division in branches that continue forming new phytomers in shrubs, trees or inflorescences (Howell, 1998; Leyser and Day, 2009).

Meristematic tissue regulates plant architecture by stem cell differentiation (Barton and Poethig, 1993). Many of the mechanisms involved have been elucidated by forward genetics, i.e

genetic screens on induced mutants, and new discoveries are still ongoing in the field. Reviews by Kozuka et al. (2005); Tsukaya (2004) and Tsukaya et al. (2002) address the genetic control of leaf and petiole development and Bar and Ori (2014) reviews the effects of hormones and transcriptional regulators in leaf initiation, morphogenesis and determination. Meristems also control the integration of environmental signals, controlling the ontogenic adaptive response to local conditions, resulting in plastic phenotypes that buffer the effect of stress and micro-environmental variation (Mandel et al., 2014; Massonnet et al., 2010; ichi Sugiyama and Gotoh, 2009; Kwiatkowska, 2008; Granier et al., 2002).

1.2 Arabidopsis thaliana architecture

Arabidopsis thaliana (Heyn.) is a model organism frequently used for genetic studies in plants. This species has a rapid development into a relative small size rosette and inflorescence allowing to grow it in small spaces and higher number of replicates than other plants (Boyes et al., 2001). *Arabidopsis* genome is relatively small (5 chromosomes with around 125 Mb) being one of the first plants in being fully sequenced (Kaul et al., 2000), evolving into a myriad of genetic resources such as inbred experimental population, e.g AMPRIL and MAGIC (Bergelson and Roux, 2010), diversity panels, e.g. 1001 Genomes Project (Weigel and Mott, 2009; Alonso-Blanco et al., 2016), mutant panels, e.g "unimutant" collection (O'Malley and Ecker, 2010; Bergelson and Roux, 2010), among many others.

Arabidopsis can grow complete its life-cycle in around 6 weeks in laboratory conditions, although some ecotypes can take several months to flower (Napp-Zinn, 1985; Shindo et al., 2007). During their vegetative phase, seedlings produce a hypocotyl with two cotyledons. The Shoot Apical Meristem (SAM) produces new leaves in a rosette until the flowering signals trigger the transition from juvenile stage to adult, reproductive stage. Arabidopsis has a determinate development, meaning, that no more rosette leaves are formed after floral initiation. The reproductive structure is a inflorescence with variable number of branches, cauline leaves and flowers (Boyes et al., 2001; Kjemtrup et al., 2003; Vanhaeren et al., 2015).

The formation of new phytomers during the rosette stage develops short internodes, resulting in new leaves growing very close to each other. This property confers the rosette structure of the species. Arabidopsis phyllotaxis initiate new leaves each 137.5°, giving a spiral-like aspect

to young rosettes (Mündermann et al., 2005; Kuhlemeier, 2007).

Camargo et al. (2014) observations on natural variation in Arabidopsis rosette development noticed heritable variation in early juvenile rosette morphology. This variation in rosette shape has been studied by Camargo et al. (2014) and Pérez-Pérez et al. (2002). Recent studies have also addressed the natural variation in *Arabidopsis* rosettes, but quantifying size rather than shape, and they are summarized in table 1.1 and reviewed by Vanhaeren et al. (2015) and Humplík et al. (2015) . Research quantifying whole-organ rosette structure, using geometrical morphometrics or statistical shape description, are absent in the literature,.and cited studies using top-view imaging of rosettes, only measure size, growth and at most the compactness, i.e. rosette coverage, but not global shape parameters.

1.3 Phenotype to Genotype Mapping in Arabidopsis thaliana

The main approaches used to unveil the genetic cause of phenotypic variation in plants are forward and reverse genetics (Alonso and Ecker, 2006). Forward genetics consist in either a search for or production of mutants which vary from the wild-type for traits of interest (Peters et al., 2003). Briefly, new mutants are produced by any method that accelerates the mutation rate in the plant, e.g by radiation or biolistics. Those mutants are then screened for phenotypes of interest, and after recognizing them, subjected to a labour-intensive procedure of searching for the specific genome region mutated, which can be cloned and inserted in a non-mutant variety for testing. Reverse genetics approach works in the opposite direction (Adams and Sekelsky, 2002), the gene or genes of interest are subjected to directed mutations, gene silencing, or other molecular techniques, to test specific relationships between the gene and the resulting phenotype. Generally, forward and reverse genetics has been able to establish the causal relationship between genotypes and resulting phenotypes.

Quantitative genetics makes use of natural variation in populations rather than mutants (Falconer and Mackay, 1996). The Quantitative Trait Loci Mapping approach consists in associating genetic variation with phenotypic variation to ascertain the genetic architecture of traits. In general, quantitative genetics provides an approximate measurement of causal loci

location in the genome and their effect in the phenotype (Holland, 2007). To identify locus location, statistical methods calculate genetic distances between phenotypes, e.g Sturtevants' linkage map on Drosophila, between a qualitative traits and a quantitative one, e.g. Sax (1923), or between a quantitative trait and genetic markers, e.g. Thoday (1961). The advent of molecular genetic markers allowed to calculate distances between them to build genetic maps (Barton and Keightley, 2002). Thus, statistical genetic tools to calculate distance between molecular markers and phenotypes opened up the possibility to map phenotypic variation to markers variation (Mackay, 2001). The rationale is that polymorphic causal loci in a population produce variation in phenotypic values in an amount determined by the loci's role in the genotype to phenotype route, which is the "effect size" of such loci, i.e the difference in alleles' phenotypic values.

The power to quantify association between causal loci and phenotypic variation through allelic variation in molecular markers is based on the linkage dissequilibrium (LD) between the loci and markers (Takuno et al., 2012). When a marker and a locus are in LD, markers' alleles act as informative proxies of allelic variation in the causal loci. The degree of linkage between phenotypic values and genotypic frequencies at any given marker allows the mapping of phenotypes to genomic regions.

However, quantitative phenotypes, those whose values are continuously distribute, are often polygenic, having several loci involved in its variation (Lynch et al., 1998). For a quantitative trait, many Quantitative Trait Loci (QTL) can be found, establishing a range of complexity in their regulation. Many situations exist, from polygenic additive traits, where many genes participates each one with small effect, oligogenic traits, where few genes determine most of phenotypic variation, to complex traits where major effect genes are interacting with each other and with minor effect genes (Collins, 2007, and figure 1.1). To unveil the genetic architecture of complex traits, it is necessary to account for as much genotypic and phenotypic natural variation as possible, and several strategies have been developed to extract the most from quantitative trait loci (QTL) mapping assays.

Experimental populations have been and are being designed and studied to improve QTL mapping in terms of power, effect size and resolution (Mitchell-Olds, 2010; Bergelson and Roux, 2010). Briefly, different experimental populations differ in their properties for QTL mapping.

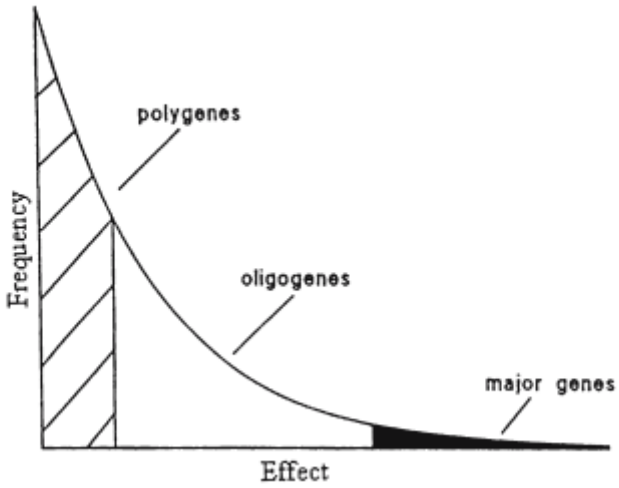


Figure 1.1: Representation of gene effects over a quantitative trait (reproduced from Collins (2007)). A quantitative trait can be affected by many genes with low effect (polygenes), or regulated by few genes (oligogenes) with intermediate effects or regulated by few genes with major effects. The categories are not disjoint, meaning a quantitative trait can be polygenic but with few genes having major effects and others at intermediate to low effects.

The key factors are the amount of genetic variation in the population, the type and number of genetic markers available to assess such genetic variation, and the possibility to calculate recombination frequencies between markers, markers-putative loci and the association between phenotypes and either markers or causal loci (Korte and Farlow, 2013; Weigel, 2011).

The next two sections of this thesis briefly introduce mapping populations that have been used in this research, and the statistical methods applied to them for genotype to phenotype mapping.

1.3.1 Experimental populations for Quantitative traits mapping

Experimental populations have been designed to study quantitative genetics of quantitative traits (see reviews by Weigel (2011); Bergelson and Roux (2010) and figure 1.2). They range in complexity, genetic diversity and population structure.

Family-based designs are biparental and multiparental crosses, with little population structure and low genetic diversity, whose genetic distances can be calculated from recombination frequencies. Tipically, 6-8 generations after crosses can be reproduced by selfing to generate "immortal" populations, also called pure lines, that do not change their genotype through further generations due to being isogenic lines, i.e genome-wide homozygous.

Population-based designs are natural diversity populations, with larger genetic diversity but

potentially important population structure. Recombination frequencies are difficult to measure (Stumpf and McVean, 2003), but genome-wide association is feasible when the marker-based genetic or physical map has been calculated in advance (Mitchell-Olds, 2010; Ambrose and Purugganan, 2013). The most relevant types of populations are briefly described in this section and reviewed at figure 1.2.

Biparental cross represents the simplest, and oldest kind of, experimental population for gene mapping, so that most methods were originally elaborated for them. In a biparental cross, two varieties differing in the trait of interest and in genotypic values for a set of molecular markers, are crossed to produce a filial generation, F_1. Parentals are ideally genome-wide homozygous inbred lines, so F1 generation is genome-wide heterozygous for all those loci that differ between parentals. Therefore, F1 genotypes and phenotypes should be similar to each other, unless any of the parentals had any segregating alleles.

Successive crosses or selfing can be done from the F_1, to yield F_2, F_3, ... generations. When crossing individuals from same generation, new combination of extant alleles at multiple segregating loci are appearing in the population. In addition, recombination events "shuffle" the genomic blocks between crossovers, i.e haplotypes, getting smaller each generation. The effect of recombination in population genetics and evolution is a difficult core concept, the reader is encouraged to see classical and fundamental research like Smith (1978),Felsenstein (1974), Crow and Kimura (1965) or Webster and Hurst (2012) and references therein. If subsequent crosses are done with a parental line, i.e. backcrosses, the population results in introgression lines with genomic pieces of a parental into the genetic background of the other. Finally, from the F1 or after several generations of intercross or backcross, four or five generations of reproduction by selfing and single seed descent, in self-compatible species, or sibling mating, in self-incompatible species, generates Recombinant Inbred Lines (RILs). RILs eventually become nearly genome-wide homozygous (although some residual heterozygosity may remain) and therefore "immortal populations", i.e their genetic is fixed assuming no mutation and no contamination. RILs' genetic stability allows for successive stages of phenotyping the same line, as well as having multiple replicates of genetically identical individuals. The methods for QTL mapping in biparental populations will be reviewed in the section 1.3.2 .

On the other hand, populations of naturally occurring varieties are also useful for mapping

purposes. In Arabidopsis, due to selfing as a common way of reproduction (Charlesworth and Vekemans, 2005; Tang et al., 2007), local varieties are identified as ecotypes, being already genome-wide homozygous and thus equivalent to RILs (Weigel, 2011). Ecotype populations have the advantage of bearing more allelic and phenotypic diversity than crosses, at the same time that the number of historical crossovers is higher than in populations derived from a single cross. A drawback when analysing ecotypes is that the crossovers location cannot be placed, at least not as easy as in controlled crosses (Stumpf and McVean, 2003). For that reason, the original methods of QTL mapping cannot be used and "genome-wide association studies" (GWAS) is the appropriate tool instead (see section 1.3.2).

Although, ecotypes are expected to have a high allelic diversity, it is also expected that most phenotypic variation is due to "common alleles", that is, several alleles are in high frequencies in the population (Gibson, 2012). However, it is also common to find "rare alleles" variation, i.e. many variants exist for a single locus in the population (Gibson, 2012). In addition, natural populations have several kinds of population structure (Astle and Balding, 2009). More specifically, population structure refers to a population having nested sub-populations with different allele frequencies due to restricted gene flow, i.e non-panmictic populations. Kinship, or family-related population structure, refers specifically to clusters of individuals having common alleles at many markers due to "Identity by Descent" (IBD), meaning that they share common parents back in the population history. Cryptic relatedness refers to the presence of close relatives even when the sample has been designed to contain non related individuals or ecotypes. According to Astle and Balding (2009), all population structure categories are based in the concept of "an unknown and unobserved pedigree" (Astle and Balding, 2009). The effect of population structure in QTL mapping is an increased rate of false positives in comparison with experimental crosses. Another kind of population structure is due to selection and drift, that make that markers and QTLs can be in linkage disequilibrium at far distances even if markers in between are not linked, sometimes can be found between markers in different chromosomes (Zhang et al., 2002). The consequence is that QTL detection power and resolution is potentially limited by population structure (Patterson et al., 2006).

Multiparental crosses are in the midpoint between biparental crosses and natural populations in several aspects (Darvasi and Soller, 1995). They are designed to have higher variation

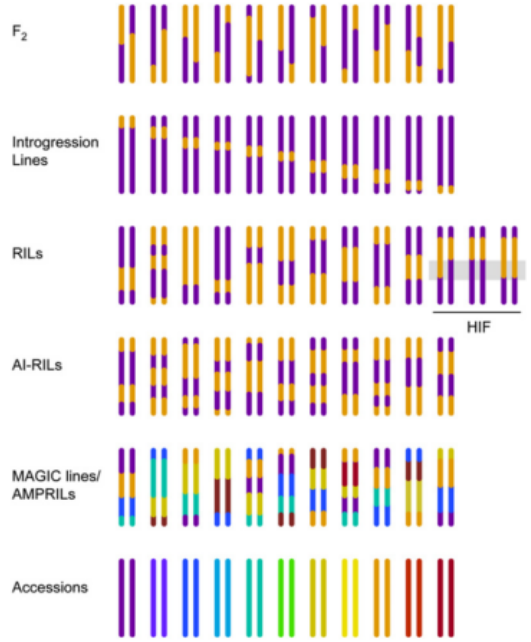
than crosses, that depends on the number of founder individuals and the natural variation between them. At the same time, parentals are crossed for several generations , so the cumulative number of crossovers breaks parental haplotypes in smaller sizes. The crosses eliminates far distance linkage disequilibrium due to selection. Some examples of multiparent crosses are: Multiparent Advanced Generation Intercross (MAGIC)(Kover et al., 2009), Advanced Intercross Recombinant Inbred Lines (AI-RIL) (Balasubramanian et al., 2009) and Arabidopsis Multiparent Recombinant Inbred lines (AMPRIL) (El-Lithy et al., 2004; Huang et al., 2011).

In multiparent crosses, genetic markers genotypes cannot be assigned immediately to a parental as in biparental populations, yet, imputation methods can be performed to assign a probability for a marker coming for a specific parental (Broman, 2012; Zheng et al., 2014). Multiparental crosses, albeit requiring complex statistical methods for QTL mapping, fix most of the drawbacks of biparental crosses and natural populations while maintaining a certain level of their advantages (Kover and Mott, 2012). Several Arabidopsis multiparental crosses are available, being the most widely used AMPRIL (El-Lithy et al., 2004) and MAGIC (Kover et al., 2009). In this thesis the MAGIC population will be used being a cross of 19 parentals, intercrossed for four generations and selfed for 6 generations (Kover et al., 2009).

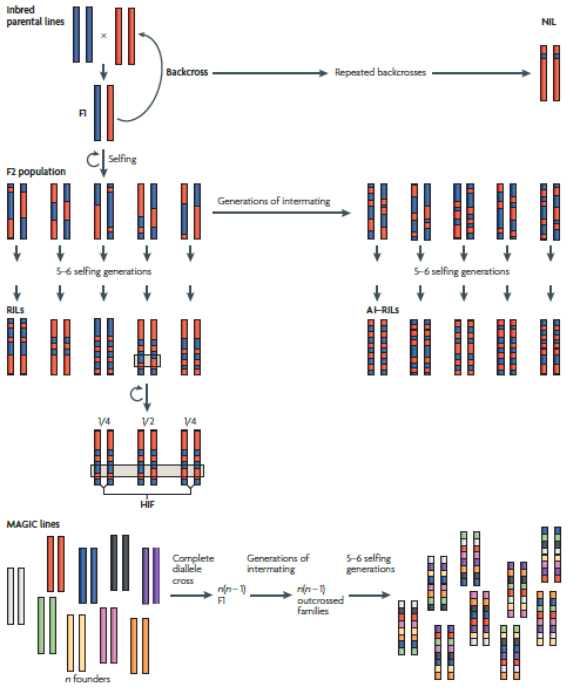
In general, biparental crosses allow an initial mapping of traits, while natural populations are capable of more precise mapping (Kover and Mott, 2012; Keurentjes et al., 2011). However, natural populations have high proportion of spurious association, i.e. false positives, that do not occur in biparental crosses. Multiparental crosses put together the best properties of both kind of populations, amount of natural variation and fine mapping, and reduce the population structure that drives to high false positive rates (Kover and Mott, 2012).

1.3.2 Quantitative Trait Loci mapping

The fundamental technique for Quantitative Trait Loci (QTL) mapping consists in the statistical association between genetic variation, i.e. alleles of a polymorphic molecular marker, to quantitative phenotypic variation. Very briefly (see Broman (2001),Zou and Zeng (2008) and Hayes (2013) for detailed reviews), a normally distributed phenotype is split in groups according to alleles in a polymorphic locus, e.g biallelic Single Nucleotide Polymorphisms (SNPs). If the loci being tested has any effect on the phenotype and different alleles in the population



(a) From (Weigel, 2011)



(b) From (Bergelson and Roux, 2010)

Figure 1.2: Schematic description, from (Weigel, 2011) and (Bergelson and Roux, 2010), of experimental populations for Quantitative Trait Loci mapping. F_2 is the result of a biparental cross into a $F1$ generation and second step of crossing F_1. Subsequent crosses, from F_2 to F_3 and so on, would accumulate recombination events while the heterozygosis would be the expected from Hardy-Weingberg equilibrium. Introgression lines result when F_1, F_2, etc. are backcrossed repeatedly with one of the original parentals. If anyF_n generation is reproduced by selfing and single seed descent, it becomes (nearly) genome-wide homozygous producing Recombinant Inbred Lines (RILs). Heterogeneous Inbred Families (HIF) are a subset of RILs with residual heterozygosity for a small region, being near isogenic lines (NILs), except for that region. They are useful for QTL confirmation and fine mapping. Advanced Intercross Lines (AIL-RILs) are originated as RILs, but instead of selfing and single seed descent from F_1 or F_2, several generations are intercrossed, generating more mosaic genomes from both parents and later are selfed to obtain RILs. Multiparent lines, like MAGIC or AMPRIL, uses the Advanced Intercross approach with multiple parents, tipically 6 or 8 but MAGIC used 19, so increase the genetic diversity at the time that build genetic mosaics of many parentals, e.g each MAGIC line has genomic pieces of 9 parentals in average. Finally, natural accessions, plants collected from their original location and environment represent the most variable set, however, recombination frequencies are not possible to calculate, so that specific methods for QTL mapping has been developed for them.

are associated to variation in the phenotype, the phenotypic values would be split in a mixture of normal distributions, whose "effect size", i.e. distance between means, is proportional to the loci effect on the phenotype (Zou and Zeng, 2008). On the contrary, if such a locus has no relationship with the phenotype, the mixture of distributions results in a low difference of means. The significance of such difference in means is measured by Likelihood Ratio or F-test p-values and indicates whether the locus is associated with the phenotype or not.

The key concept in QTL mapping is linkage (Falconer and Mackay, 1996; Lynch et al., 1998; Astle and Balding, 2009, and figure 1.3). Two genes, markers or any other DNA elements are linked when they are close in the genome, so few recombination events occurs between them. On one hand, if two elements are linked, the recombination rate between them translated to genetic distance in centiMorgans (cM) as 1 cM = 1% recombination rate (Kosambi, 1943). On the other hand, in a population and after estimating the recombination rate between flanking markers, the probability of an allele in a close locus can be calculated from the allele at the linked marker (Broman, 2012). However, even if the recombination rate cannot be calculated, as in natural ecotypes, a marker can still act as a proxy for phenotypic effects of linked locus (Astle and Balding, 2009).

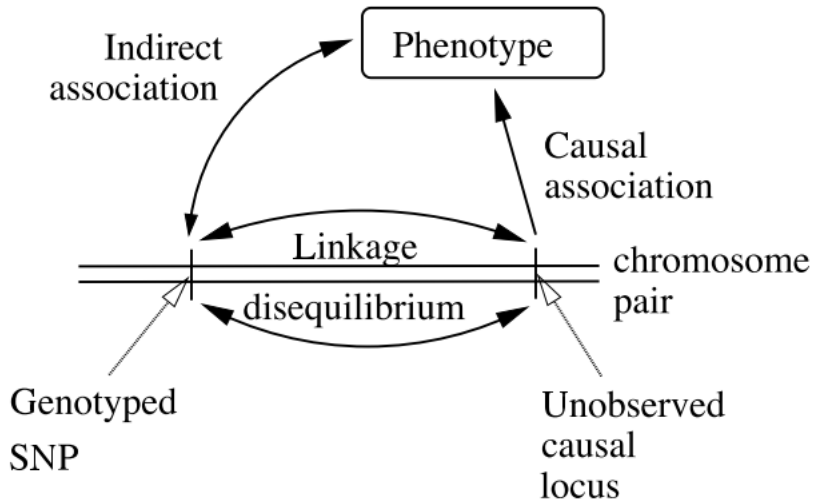


Figure 1.3: Reproduced from (Astle and Balding, 2009). Diagram representing the linkage and linkage dissequilibrium between genetic elements, e.g genes and markers. Linkage Dissequibrium results in genetic correlation between markers and gene-markers genotypes, so acting as proxies of each other. The statistical association between a non-causal marker and a phenotype can represent the causal action of a genetic element nearby.

The nature of experimental populations impose the use of different techniques for QTL map-

ping purposes Weigel (2011); Bergelson and Roux (2010), and the most common are described hereinafter.

Biparental Cross population

In biparental cross populations the hypothesis tested is whether a Quantitative Trait Locus is present and linked to a testing marker in the population under analysis (Haley and Knott, 1992). Genotyped markers can be scanned individually (Haley and Knott, 1992; Lander and Botstein, 1989), or can be the scaffold to build pseudo-markers, i.e artificial markers genotyped according to the recombination frequency between neighbouring markers (Knapp et al., 1990; Knapp and Bridges, 1990). These pseudo-markers allow to apply Interval Mapping tools (IM) (Jansen and Stam, 1994). In Interval Mapping, several intervals can be tested simultaneously by a technique called Composite Interval Mapping (CIM) (Zeng, 1994; Jansen and Stam, 1994). Similarly, complex QTL models can be fitted using several markers, or pseudo-markers, simultaneously as cofactors, the technique is called then Multiple QTL Mapping (MQM) (Zou and Zeng, 2008). The core of these techniques is that recombination rates can be calculated, and the probability for the genotype of unobserved markers can be estimated (Broman, 2001), given that a markers genetic map has been built in advance.

The significance of a QTL model is assessed by a threshold on the p-value or "Log of Odds" score (LOD score) for the test (Van Ooijen, 1999). Generally either a fixed LOD score is chosen as threshold, or a permutation test is performed (Churchill and Doerge, 1994; Doerge and Churchill, 1996).

Natural Ecotypes population

As it has been stated, QTL mapping methods for biparental crosses strongly rely in the ability of calculating recombination frequencies (Doerge, 2002). This approach is not possible in natural populations, where alleles from many origins have been mixed for many generations (Balding, 2006). The approach generally taken in such populations is the so called Genome-Wide Association Studies or GWAS (Bush and Moore, 2012). The rationale behind GWAS is that it exploits the linkage disequilibrium between markers and causal QTLs (Astle and Balding, 2009, and figure 1.3). For this reason, GWAS is also called Linkage Disequilibrium mapping. The main difference is that models cannot be weighed by recombination probabilities

between markers, so that only actual markers, rather than pseudo-markers, can be tested as proxies for causal QTLs by means of Analysis of Variance test or non-parametrical versions of it (Hayes, 2013; Balding, 2006; Lazzeroni, 2001). A test is calculated to each marker, i.e. "a genomic scan", resulting in a set of p-values for each one. Since markers are arranged in a map, p-values also do it, and plots, called Manhattan plots, assist visually to detect peaks close to potential QTLs. To accept a peak as a potential QTL is not enough to assume a 5% significance, so corrections such Bonferroni correction (Korte and Farlow, 2013), False Discovery Rate correction (Benjamini and Hochberg, 1995) or Genomic inflation factor (Yang et al., 2011) have to be applied due to multiple testing. These corrections are needed to avoid spurious false positives and they are calculated on the p-values resulting after test. In addition, population structure correction and other factors can be added to the models to remove their effect during the modelling stage.

Multiparent Crosses Population

In multiparent crosses different strategies can be used according to peculiarities in each specific cross (Darvasi and Soller, 1995; Cavanagh et al., 2008). For Arabidopsis AMPRIL population, Huang et al. (2011) chose 8 founders from different regions and crossed them pairwise, thus obtaining 4 two-way hybrids. Diallel crosses between these hybrids resulted in 6 four-way "F1 crosses" (Huang et al., 2011). Each cross were selfed to F5 resulting in 532 lines.

For Arabidopsis MAGIC population (Kover et al., 2009), the strategy is intermediate between biparental and natural populations. The 19 parentals (from Gan et al. (2011)) were crossed in a way that parents and "daughters" could not be tracked to calculate Identity by Descent probabilities. However, a "Hidden Markov Model" algorithm was used to find the most probable parental of origin for each marker in each individual, based on neighbouring markers (Mott et al., 2000; Zheng et al., 2014). At the end of the algorithm, the markers has the name of the parental-of-origin as alleles for each individual. Genomic scan test each marker sequentially for phenotype to genotype association, being a GWAS-like strategy. However, instead of grouping the population in two groups, one for each SNP allele, it is split in 19 groups, one for each parental of origin, and one-way ANOVA is performed. In such way, the parental with highest influence in the phenotype can be found in *a posteriori* analysis.

Mitchell-Olds (2010) review a comparison of methods and results obtained by Atwell et al. (2010), using GWAS in a natural population, and Kover and Mott (2012), using MAGIC, for history-life traits. Other publications have explored the use of multiple populations for natural variation QTL mapping in Arabidopsis, e.g Simon et al. (2008) (5 RILs populations with consensus SNPs for flowering time, rosette size and fitness as seed production), O'Neill et al. (2008) (6 RILs populations with genetic maps aligned to markers physical position for flowering time and seed lipids). Brachi et al. (2010) studied a comparison of QTL mapping for flowering time using a set of biparental populations and GWAS on natural population.

1.4 High throughput phenotyping

Plant characterization in terms of morphometric, physiological or molecular phenotypes has traditionally required many manual procedures that are labour intensive, time-consuming and involve destructive sampling of individuals and organs. With the development of genotyping and gene transformation technologies, phenotyping became a "bottleneck" that delays the performance of screening and gene discovery studies. However, techniques that utilize cameras and sensors to measure the physiological status, the morphometric variables and several other traits have been developing during the last decade (Houle et al., 2010; Furbank and Tester, 2011; Pieruschka and Poorter, 2012; Rahaman et al., 2015).

Those tools are being mounted on robotic devices able to handle individual plants, or trays, along their life cycle, i.e "from seed to seed", under controlled greenhouse conditions (Dhondt et al., 2013; Brown et al., 2014; Granier and Vile, 2014). The results are phenotyping devices able to support large plant populations, at controlled conditions, and tracking plant features dynamically, i.e. individual time series, non-destructively and in a high-throughput manner (Sozzani et al., 2014).

The kind of sensors in use in phenomics, the art of high-throughput phenotyping, are mainly imaging sensors, i.e. chips containing an array of light sensors able to digitally image a scene in one or several wavelengths (Granier et al., 2006; Jansen et al., 2009; Arvidsson et al., 2011; Zhang et al., 2012; Tisne et al., 2013; Dhondt et al., 2014). Examples are fluorescence, Far-Infrared (FIR), Near-Infrared (NIR), multi-spectral and hyper-spectral cameras, and the traditional visible spectra cameras (RGB from red-green-blue). Each wavelength is potentially able

to describe one or several properties of plants. Fluorescence, together with so-called actinic light, i.e. photosynthetically active, and flash pulses provides information of photosynthesis related pathways. Far-infrared light emission by plant is a measurement of temperature that may be correlated with drought and salt stress, as well as biotic damage by pathogens Humplík et al. (2015); Rahaman et al. (2015); Rahman et al. (2015).

From RGB measurement two sources of information can be obtained. On one hand, colour is collected from light intensity at visible spectra coded as the combination of red, green and blue intensities. Colour helps to investigate aspects of plant physiology since many processes have marked colour changes, e.g pathogen attack on leaves turns in colourful structures with identifiable shape, chlorosis or senescence (Mutka and Bart, 2015). On the other hand, RGB images allow studying plant mophological structure at different levels (Furbank and Tester, 2011; Brown et al., 2014). There are studies where the whole plant structure is studied, or just part of them like roots, leaves, rosettes (Dhondt et al., 2013; Chen et al., 2014; Granier and Vile, 2014; Sozzani et al., 2014). Plant images permit the measurement of morphological traits in the pictures generated. The advantage is that pictures can be saved for later analysis. Measuring leaves or any other part of the plant manually is a slow and labour-intensive, and sometimes destructive, procedure. Therefore, the possibility of measuring leaves or rosettes using images, allows an increase in the number of samples, the use of non-destructive procedures, and calculation of novel metrics that were not pre-planned or that only can be obtained from digital (quantized) objects . In addition, pictures can be taken at several time points, and the range depends on the automation degree, e.g. FIR images can be taken at a range of seconds or minutes, while RGB pictures of many plants can be taken everyday or several times per day.

1.4.1 Computer Vision - Image Analysis of Arabidopsis rosettes

The automation of plant imaging for phenotyping typically results in huge amount of images. It is possible to take manual measurements on a selection of them, but to extract most of their information is intractable by non-automated methods. Image Analysis by methods of Computer Vision allows quantification of traits of interest in a automatic or semi-automatic manner. For Arabidopsis rosettes, at the time of this project was started, several software tools were available as sub-products of research projects and commercial phenotyping platforms.

Initially, plants need to be extracted from pictures as image objects, a process called segmentation. As examples, Lemnatec High Througput Phenotyping (HTP) device includes the software LemnaGrid that allows users to implement image analysis pipelines for segmentation of plants. Klukas et al. (2012) implemented a set of image processing pipelines for Arabidopsis rosettes, maize and other crop shots, called Integrated Analysis Platform (IAP) that extends and organizes LemnaGrid methods. The Donald Danforth Plant Science Center has adapted the library OpenCV (Bradski, 2000) to plant computer vision purposes, with special emphasis in Lemnatec derived images, calling it PlantCV Fahlgren et al. (2015a). Other plant image processing tools are HTPheno (Hartmann et al., 2011), TraitCapture(Brown et al., 2014, work in progress) and Phenotiki (Minervini et al., 2017) among others. Finally, some software tools dedicated to plant rosettes are Rosette Tracker (De Vylder et al., 2012) and Phytotyping4d Apelt et al. (2015). Rosette tracker is a tool for rosette segmentation and analysis of colour, fluorescence and temperature. It describes rosettes shapes by using their Area, Diameter, Stockiness (roundness), Relative Growth Rate and Compactness. Phytotyping4d is optimized for 3D light-field cameras (Raytrix GmbH) allowing to separate and identify leaves in the image, allowing more specific set of parameters to study.

General purpose software is also extensively used in plant phenotyping, Matlab (MathWorks, 2009) and ImageJ (Abràmoff et al., 2004) being the most popular. These programs have the advantage of allowing the design of specific scripts and pipelines for specific questions, as well as design or application different methods to study different (computational) problems in plant biology, e.g counting leaves, measuring petiole length, identifying leaves tips, generate complex shape metrics as “shape context”, angles, etc..

Recently, some effort has been done in using RGB cameras to describe leaves without cutting them off the plant. This requires the segmentation of the rosette, and later to split the rosette in leaves, even if they are overlapping. This task is not easy, and international collaborations are still ongoing, like the Leaf Segmentation Challenge at the Computer Vision Problems in Plant Phenotyping conference. Recent methods in this field are Minervini et al. (2014); Pape and Klukas (2014); Giuffrida et al. (2015).

1.4.2 Morphology measurements

Morphometry is a key topic in biology, since shape is a significant factor to study how species adapt, evolve, function in their environments and also for biological classification and taxonomy. However, a biological definition of shape is still absent in the literature. Mathematicians define shape as the geometrical information that remains when location, rotation and scaling is removed (Kendall, 1984; Claude, 2008). Under this definition, two shapes are equivalent if a set of translations, rotations and scaling operations can be found so that one shape can be transformed in the other. However, the latter definition is not enough for non-rigid, deformable objects whose shapes being alike cannot be found equivalent by rigid transformations. The formal study of such non-rigid objects and its properties lays on the realm of mathematical topology (Mardešic and Segal, 1982). Fortunately, other methods have been built to compare shapes of objects by the incorporation of statistics.

Generally, evolutionary, developmental biologists and paleontologists, among others, utilize the tools of geometrical morphometrics to formally compare biological structures that deviates through time by different growth rates, in what is called allometric growth (Zelditch et al., 2012; Small, 2012, for review of methods). The foundations of Geometrical Morphometrics were initially proposed by D'Arcy Thompson (Thompson et al., 1942), and developed through the work of Bookstein (landmarks),Rohlf (biometrics), Kendall (shape spaces), Klingenberg (fluctuating symetry, evo-devo and QTL for shape) and many others along XX century (Zelditch et al., 2012) . The power of Geometrical Morphometrics relies in the definition of landmarks, homologous points in the classical biological sense, that can be found in all the specimens in the sample. Distance functions to measure shape similarity between objects and perform statistical comparisons have been developed such as the Procustes distance, i.e. a metric on distance between landmarks, among others. The use of images is fundamental in geometrical morphometrics, to be able to compute positions in a common framework and distances, so software to analyse such images, by manually placing landmarks has been developed, like MorphoJ (Klingenberg, 2011), TPS (Rohlf, 2015) and IMP (Sheets, 2003). Automatic placement of landmarks (Iwata, 2012) is only feasible in a few specimens and controlled conditions, using the so called pseudo landmarks. The ability of finding landmarks automatically in plants would represent a enormous improvement in automatic phenotyping of organisms.

For objects without clear criteria to assign such landmarks, like plants in general, some other methods have been developed by modelling the outline of objects, such as SHAPE software (Iwata and Ukai, 2002), e.g. elliptic Fourier descriptors for biological shapes such as petals (Yoshioka et al., 2004) or leaves and leaflets (Chitwood et al., 2012, 2013a).

Another set of methods to describe and calculate shapes is digital geometry. In 2D and 3D digital geometry, objects' geometrical properties are modelled as a set of features. A type of those features are shape descriptors that either measure the departure from an expected geometry, e.g roundness as departure of a circle, or condense the spatial structure of an object into scalar or vector-valued metrics (Nixon and Aguado, 2012, chapter 7;Zhang and Lu, 2004). Such methodology has been used mostly for object retrieval in computer vision with the objective of recognizing objects in a image by comparison with an array of structural properties of such objects in other images after a statistical learning process. The idea is to generate a statistical multivariate model that defines a certain object by their shape metrics, accepting that not all the information can be capture by a single metric, and also that any of those descriptors provide an accurate measurement. However, the joint and conditional distributions of the multivariate feature space has proved useful for this image retrieval.

Camargo et al. (2014) proposed a set of shape descriptors to analyse Arabidopsis rosette morphology. These are global, whole-organ, measurements based on 2D distribution of pixels. These descriptors were originally implemented in the LemnaGrid software, but their work opened the formulas and algorithms for being used independently of such platform. The approach in this thesis is to use the descriptors proposed by Camargo et al. (2014) as rosette morphology measurements. In the appendix A, a succinct description of the shape descriptors use in this thesis is provided together with their rosette descriptive value and a short introduction to digital imaging.

1.5 Thesis purpose, background and structure

Observations on natural variation of Arabidopsis thaliana juvenile plants, from seedlings to flowering, suggest that each ecotype has a defined rosette developmental pathway. Although the rosette habit of different ecotypes has a different aspect, it remains similar between plants from the same ecotype. This suggests genetic control of the developmental route during the

juvenile establishment. Rosette aspect differs generally in the size of leaves, showing differential allometric growth between several parts of the leaf. As examples, some accessions display long leaf blades, by longitudinal growth rather than sagittal, while others are near isometric, i.e rounded. Some accessions have long petioles, connected either to long or short blades, that can be measured by the blade length over petiole length ratio. Leaves can have serrate margins, curvatures, etc.

Current research on leaves shape has found many genetic determinants for traits like leaf roundness, blade/petiole lengths, leaf border complexity (Bar and Ori, 2014; Kozuka et al., 2005; Tsukaya, 2004). However, no studies, to my knowledge, have made extensive use of overall rosette description, although some studies has used whole-rosette measurements of projected rosette area and compactness. Humplík et al. (2015); Vanhaeren et al. (2015) and table 1.1 review recent works on Arabidopsis rosettes phenotyping. On the other hand, it is well known that leaves development reacts in response to environmental variation in light, temperature and others, e.g. Cookson et al. (2006); Hopkins et al. (2008); Mishra et al. (2012); Chitwood et al. (2016) and a review by Tsukaya (2004), and it is expected that leaves and rosettes phenotypic variation is affected by genetics x environment interactions.

In this thesis, I will explore natural variation in Arabidopsis thaliana rosette morphology by quantifying shape descriptors from digital imaging. The research questions are whether rosette morphology exhibit genetic variance and environmental variance, and the relative role of both in the phenotype. Quantitative trait loci mapping will potentially help to determine the genetic architecture of such complex trait, and generate hypotheses on potential candidate loci coordinating genetics and environment effect for future research.

The thesis structure consist in three chapters with the results from the study of a natural accessions population by Genome-Wide Association mapping (chapter 2), a Recombinant Inbred Lines from a biparental cross (Cape Verde Island x Argentat) by Linkage mapping (chapter 3) and a Multiparent Advanced Generation Intercross (MAGIC) by association mapping (Chapter 4). A general discussion gathering common results and future prospect is provided at Chapter 5.

Paper	Goal	Methods	Measurements
Leister et al. (1999)	Biomass and growth estimation	Home made automatic phenotyping. Time lapse.	Rosette Area. Biomass-Area correlation. RGR
Pérez-Pérez et al. (2002)	Natural variation in rosette architecture. QTL mapping.	Manual phenotyping.	Leaf area, perimeter, length and width. Marginal serration. Leaf number. Major and minor rosette diameters. Bushy/Loose rosettes
Chenu et al. (2004, 2005)	Rosette architecture in reduced light intensity	Manual phenotyping.	Leaf growth. Petiole and blade morphology. Leaf initiation, expansion rate. Phyllotaxic and zenithal angles.
El-Lithy et al. (2004)	QTL mapping growth.	Manual phenotypping. Time-lapse.	Rosette area. RGR. Specific leaf area (SLA).
Granier et al. (2006)	Sensitivity to water deficit	PHENOPSIS. Drought stress Time-lapse	Leaf number, rosette area, leaf area.
Walter et al. (2007)	Seedling growth acclimation	GROWSCREEN. Time Lapse.	Rosette area. RGR.
Jansen et al. (2009)	Stress Tolerance	GROWSCREEN FLUORO. Time-lapse	Rosette area, RGR, compactness, stockiness
Pérez-pérez et al. (2011)	Multilevel trait correlation	Home-made automatic phenotyping. Rosettes, leaves and cells	Rosette area, perimeter, maximum and minimum Feret's diameter. Best fitting ellipse area, perimeter, major and minor axis. Rosette compactness and evenness.
Arvidsson et al. (2011)	Genotype effects	Scanalyzer. Home-made imaging system. Time-lapse	Rosette area, convex hull, compactness. RGR
Zhang et al. (2012)	Natural variation of growth and development	Home-made automatic phenotyping. Time-lapse	Rosette area, Compactness. RGR.
De Vylder et al. (2012)	Software development	Rosette Tracker.	Area, Feret's diameter, stockiness, compactness. RGR
Green et al. (2012)	Quantification of 2D phenotypes from images	Phenophyte. Time-lapse. Herbivory. Phytopatology	Rosette area, leaf area. RGR
Klukas et al. (2012)	Software Development	Integrated Analysis Platform.	Plant size. Skeletons. Convex Hull. Leaf Tips.
Tessmer et al. (2013)	High throughput growth Analysis	HPGA. Time Lapse	Plant Area. Absolute Growth Rate. RGR. Non-linear Growth Models (ODE) parameters.
Tisne et al. (2013)	Environmentally Robust Rosette Analysis	Phenoscope. Time-lapse.	Rosette area. Radius circle encompassing rosette.
Dhondt et al. (2014)	Time-Resolve rosette growth	In Vitro Growth Imaging System. Time-lapse. Salt, Osmotic and Oxidative Stress	Rosette Area, Compactness, Stockiness. RGR
Wilson-Sánchez et al. (2014)	PhenoLeaf database	Manual phenotyping	Rosette Area. Compactness (as Area/fitting ellipse Area). Leaf Serration. Leaf colour. Lamina shape, symmetry, relative size. Leaf Number. Phyllotaxis. Margin shape. Petiole length and width.
Bac-Molenaar et al. (2015)	GWAS drought stress	PHENOPSIS. Time-lapse. Drought stress	Rosette area, exponential growth.
Apelt et al. (2015)	Spatio-temporal plant growth	Phytotyping4d, Light field camera.	Leaf appearance, hyponastyc movement, leave and rosette shape. Leaf size, compactness,circularity, aspect ratio, curvature, leaf angles. Rosette circularity, compactness, touch, asymmetry, reliability angle. Relative expansion rate.
Clauw et al. (2015)	Leaf response to drought stress	WIWAM. Genome-wide transcriptome analysis	Rosette Area
Kooke et al. (2015)	QTL mapping phenotypic plasticity	Manual phenotyping.	Leaf Area. RGR. Rosette branching. Internode length. Plant height at 1^{st} silique
Kooke et al. (2016)	GWAS Morphological traits	Manual phenotyping.	Leaf Length, Petiole/blade length ratio. RGR. Rosette branching. Total plant height. Plant height at 1st silique
Flood et al. (2016)	Fluctuation in heritability	Phenovator. Time-lapse.	Rosette area
Viaud et al. (2017)	Organ-scale plant model	Phenoscope. Time Lapse. Functional-structural modelling	Leaf distance, angle and area. Phyllotaxis
Minervini et al. (2017)	Hardware Development	Phenotiki. Cheap phenotyping platform. Analysis by iPlant/CyVerse BisQue platform	Rosette area, Diameter, Perimeter, Convex Hull Area. Compactness, Stockiness, Leaf Count. RGR

Table 1.1: Studies on Automatic Phenotyping of Arabidopsis Rosettes

Chapter 2

Genome-wide Association of Rosette Structure in a Natural Ecotypes Population

2.1 Introduction

This chapter describes the genetic architecture of Arabidopsis thaliana rosettes morphological traits as revealed by Genome-Wide Association mapping performed on a population of Natural Ecotypes from a diverse range of North Hemisphere locations.

2.1.1 Genome-wide Association Studies - Generalities

Genome-wide association studies (GWAS) are a set of statistical genetics methods for relating phenotypic variation to genotypic variation (Ingvarsson and Street, 2010; Ogura and Busch, 2015), whose ultimate goal is to elucidate the genetic architecture of complex traits (Bush and Moore, 2012; Mitchell-Olds, 2010; Weigel, 2005). The technique requires a population that exhibit extensive natural variation in the trait under study (Alonso-Blanco et al., 2009), and it is designed to cope with individuals whose family structure has not been tracked, as opposite to experimental cross and pedigrees. Population genetic diversity is captured by molecular markers, usually Single Nucleotide Polymorphisms (SNPs), previously mapped in the chromosomes (Hayward et al., 2014).

GWAS tests for statistical association between molecular markers alleles and phenotypic values, assuming that molecular markers are in Linkage Dissequilibrium with causal loci in their vicinity (Astle and Balding, 2009; Brachi et al., 2011; Korte and Farlow, 2013, and figure 1.3). Linkage Disequilibrium (LD) refers to the phenomena of two loci having joint frequencies, in a given population, that departs from expected frequencies in Hardy-Weinberg equilibrium (Gupta et al., 2005; Falconer and Mackay, 1996). Linkage Dissequilibrium is measured as a correlation-like coefficient between markers, equivalent to the common information shared between them (Lazzeroni, 2001; Devlin and Risch, 1995). Thus, GWAS consists of test loci across the chromosome, hence genome-wide, through the molecular markers genotyped nearby. The power to detect loci-phenotype dependence through marker-phenotype association depends on the extent of Linkage Dissequilibrium (Vos et al., 2016). LD decays with distance respect of a given locus to "background levels". If LD decay is low, therefore loci that are far remain in LD, a causal locus is in LD with distant markers and they will be associated to the phenotype, hindering the mapping resolution. On the other hand, if LD decay fast, haplotypes, i.e. genomic blocks transmitted as a unit through generations in a population, become shorter and mapping resolution is improved (Buckler and Gore, 2007). The optimum number of molecular markers required to interrogate the whole genome at the resolution dictated by LD decay is specific for each population, as it is for LD and LD decay (Korte and Farlow, 2013; Zhou et al., 2012).

In a experimental population whose individuals come from natural populations rather than crosses, a certain degree of population structure exists. It may form clusters of closely related individuals with recent common origin, so that few crossover has occurred in their chromosomes making larger haplotypes and lower LD decay (Astle and Balding, 2009; Hayes, 2013). This effect of population structure is related to inbreeding (Slatkin, 2008), and may be important in high-selfing species, like Arabidopsis (Ohta, 1982; Long et al., 2013; Buckler and Gore, 2007). In addition, selection can favour LD between loci that are far apart, even in different chromosomes, if they area advantageous in certain environment (Nielsen, 2005). In this case, called phenotypic correlation (Searle, 1961), false positive associations are found since a locus for a correlated trait shows as a potential QTL for the phenotype under study. In populations that has suffered bottlenecks, genetic drifts may result in the similar issues.

2.1.2 Genome-wide Association Studies - Calculations

The statistical procedure in GWAS is a whole-genome scan, marker by marker, applying an hypothesis test, e.g. student T-test, Wilcoxon test, Analysis of Variance or Linear Mixed Models, and recording the probability value, i.e. p-value, of significance Balding (2006); Hayes (2013). The use of Linear Mixed Models allows one to accommodate co-variables and variance-covariance error structure in the models that account for population structure, preventing the issues aforementioned (Zhang et al., 2010; Lipka et al., 2012; Tang et al., 2016). Also several genetic models are available such as additive and dominant models. Population structure is separated in a kinship matrix (VanRaden (2008) and Zhao et al. (2007, with a focus in the Arabidopsis thaliana experimental population used in this chapter)), accounting for family-based substructure (Astle and Balding, 2009), and a Principal Component Analysis-based set of variables accounting for markers variation Patterson et al. (2006); Price et al. (2006, 2010). Others methods to account for population structure are available such as the software STRUCTURE and its derivates (for a review, see Novembre, 2016, and references therein).

Markers significance values, p-values, are arranged by physical position of markers along chromosomes in the so-called Manhattan plots. Generally, p-values are kept under a baseline, while makers associated with the traits are arranged as peaks. According to LD decay, trait-associated markers are sorted in peaks as broad as the number of markers in LD with potential causal loci.

P-values need to be corrected for multiple hypothesis testing, so that the number of false positives are reduced. The usual methods are the Bonferroni correction, considered too stringent, and the False Discovery Rate (Benjamini and Hochberg, 1995).

2.1.3 Natural Ecotypes Population

Atwell et al. (2010) demonstrated the proof-of-concept of GWAS in Arabidopsis for 107 phenotypes, mostly history-life traits, e.g flowering time, and resistance to pathogens (Mitchell-Olds, 2010; Korte and Farlow, 2013). It can be claimed that Atwell's population had been designed to have enough genetic diversity and an appropriate number of genetic markers to successfully perform GWAS. In addition, the population has been sampled from the whole range of natural location of Arabidopsis in the North Hemisphere. For this reason, this population

was chosen for an initial approach to determining the genetic architecture controlling rosettes structure. It is expected that this population shows a representation of the phenotypic space for rosette structure, as well as variation in relevant loci that control the traits. The history of the population's development is briefly presented together with relevant information about its characteristics.

The development of this population has been summarized at Gregor Mendel Institute Github wiki (Seren, a). It was initiated by Nordborg et al. (2005) for the 1001 genomes initiative (Weigel and Mott, 2009) and was originated in two separate laboratories. At first, 96 accessions were sampled in a collaborative effort by the laboratories of Bergelson, Kreitman and Nordborg (Seren, b) with the purpose of genotyping the accessions, evaluate the potential for studying population structure and linkage dissequilibrium and the feasibility of performing Genome Wide Association mapping in Arabidopsis. Nordborg et al. (2005) revealed extensive population structure and isolation by distance in this population. Some subpopulations seems to be mixed, so that some alleles are shared between far-located populations. Nordborg et al. (2002, 2005) found that Linkage Dissequlibrium in this population fall faster than expected, around 25kb, so that marker density for GWAS studies should lie under 1 markers every 100kb (Aranzana et al., 2005; Zhao et al., 2007). Authors concluded that for effective GWAS studies, the population should rise around 1000 strains and 250000 SNPs, according to Kim et al. (2007) simulation-based studies on tag-SNPs, so the number of false positive due to confounding population structure is reduced (Zhao et al., 2007).

From these 96 accessions, the laboratories of Ecker and Weigel chose a set of "20 maximally diverse" lines and re-sequenced with Perlegen technologies(Clark et al., 2007). Finally, Bergelson, Nordborg and Borevitz laboratories joined to generate a 250k Affymetrix genotyping chip (Kim et al., 2007) that was used to continue genotyping 1307 accessions (Horton et al., 2012) as the most recent resource in Alonso-Blanco et al. (2016). All available information about these genotypes can be found at the web portal of "1001 Genomes Project" [1].

This population has been used for GWAS studies by Atwell et al. (2010) and Li et al. (2010) . Atwell's studied 107 phenotypes on 199 accessions, 96 from Nordborg et al. (2005) (the population used here) and 94 from Brachi et al. (2010) . Li's paper studied flowering time under various climate conditions in 473 partially in common with Brachi et al. (2010) and

[1] `http://1001genomes.org/`

Atwell et al. (2010). Atwell and colleages measured 107 phenotypes related with plant growth, resistance to pathogens, fruit size and shape, seed dormancy, etc. Their GWAS study found "many common alleles of major effect" (Atwell et al., 2010) and candidate genes, selected *a-priori*, were significantly over-represented in the associations found, together with others whose interpretation as false or true association need to be validated, but authors realized that complex genetics and population structure can influence as confounding factors. This paper can be interpreted as a "proof of concept" that GWAS is feasible to search for candidate genes in Arabidopsis (Atwell et al., 2010; Mitchell-Olds, 2010). The paper of Li et al. (2010), focus on the dissection of natural variation in flowering time in 473 accessions planted in two greenhouse simulated local climates from Sweden and Spain. Their results mapped to 12 Flowering Time (FT) candidate genes, 4 of them located in the vicinity of known FT genes, and 8 novel loci. Four out of these 12 genes correlated with the latitude of origin of the accession, and allowed to study the fitness of optimal FT according to each season and location. Interestingly, most QTL papers does not fully consider the interaction between environment and genetics and the phenotypic plasticity of physiological, life-history and morphological traits. Atwell and Nordborg (Sasaki et al., 2015) studied the effect of genetic and environment interaction in 173 natural inbred lines, showing that temperature interaction with FT is concentrated in a 0.5% of Arabidopsis genome, occurring in the 26% of the accessions. Kooke et al. (2015) found that there is variation in plasticity itself between accessions, with QTLs for morphology and plasticity overlapping, suggesting epistasis. A possible explanation for the variation in plasticity in some accessions is that epigenetics, through DNA methylation, can buffer or enhance such plasticity (Kooke et al., 2015) .

Nordborg's population was chosen for a initial QTL scan through GWAS, given the positive results achieved in previous publications.

2.2 Material and Methods

2.2.1 Population

The experiment used 94 of these globally distributed accessions (Map in figure 2.1). The table 2.2 presents accessions names, sampling locations, collector and NASC Number for the ecotypes

used in this experiment. The SNPs genotyped for the 199 accessions, therefore including Nordborg's 96, can be obtained at Gregor Mendel Institute web [2], which is a link to the genotypes used in Suzana Atwell's publication (Atwell et al., 2010).

2.2.2 Experimental set-up

Natural Ecotypes seeds were bought at Nottigham Arabidopsis Stock Centre (NASC) and bulked up in the Aberystwyth Botany Gardens. Forty to fifty seeds of each accession seeds were sowed in single moist pots and kept in vernalization at 4°C in a dark room for two days. Pots were maintained for a week in a growth chamber to allow germination and then, seedlings were pricked to single pots and moved to Lemnatec automatic phenotyping device. Pots were conical of 6 cm diameter upper base and 5.50 cm height. They were filled with 60 grams of Lenvington F2 + 20% grit/sand.

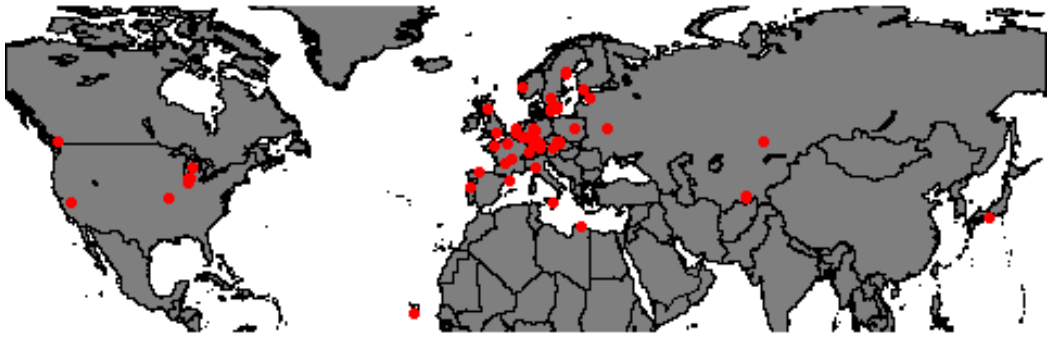
Growth chambers were set to 8 hours of light and 16 hours of dark; the greenhouse where Lemnatec device is located, is under natural daylight conditions (around 9 to 10 hours daylight in February 2015). Temperature in growth chambers and greenhouses were around 22°C during light time and 20°C at night.

Five replicates from these 91 accessions were split in five randomised blocks, one replicate from each accession in each block. Each block had 19 trays of 3x2 holes, where five pots were placed and one empty for watering the tray floor. The device automatically pour water through the hole up to a pre-set weight corresponding to an 80% field capacity.

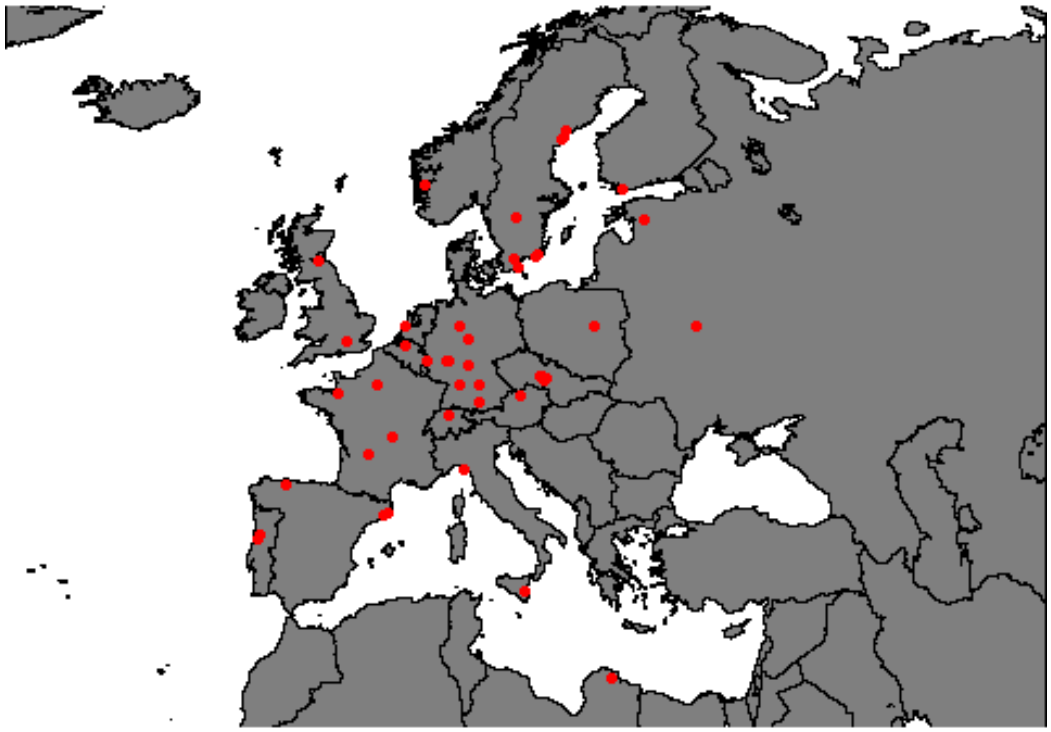
2.2.3 Rosette structure phenotyping

Rosette morphological traits were calculated as Shape Descriptor from top-view digital images.Briefly, 2D zenithal digital images were taking from Arabidopsis seedlings rosette for 13 days (DAE or Days After Experiment start). Traits were organized as 'Shape Descriptor _ DAE' as columns in data matrix, e.g Area on 2nd day as Area_2. Each row contains per-accession mean values for each trait_DAE, since most of GWAS software does not accept repeated measurements on a genotype. Experiment lasted 13 days and 16 shape descriptors were calculated as in Camargo et al. (2014) (see table 2.1 and appendix A).

[2] `https://github.com/Gregor-Mendel-Institute/atpolydb/blob/master/250k_snp_data/call_method_32.tar.gz`



(a) World wide Location of 94 Natural Accessions



(b) Location of European 94 Natural Accessions

Figure 2.1: Geographical Location of the experimental population. 94 Accessions from North Hemisphere and Cape Verde Island. Figure b focuses on European Accessions for better view. *Accessions Kas-2,CS22491 had no coordinates recorded.*

Figure 2.2: Example of Lemnatec 3x2 pots Trays. Top view image covering tray, holders and conveyor belts. The GWAS experiment was performed in the Lemnatec device, using 5 plants per tray. The pot hole at top-center position was left for automatic watering purpose, so all plants are watered from the bottom.

1. Projected Rosette Area - Area
2. Circumference - Perimeter
3. Boundary Points Count
4. Convex Hull Area
5. Convex Hull Circumference
6. Minimum Enclosing Circle Radius
7. Minimum Rectangle Area
8. Maximum Diameter - Feret Diameter
9. Compactness - $\frac{Area}{ConvexHullArea}$
10. Roundness - $\frac{Circumference^2}{Area}$
11. Boundary Point Roundness - $\frac{BoundaryPointsCount^2}{Area}$
12. Boundary Point Count to Area Ratio - $\frac{BoundaryPointsCount}{Area}$
13. Convex Hull Roundness - $\frac{ConvexHullPerimeter^2}{Area}$
14. Eccentricity
15. Normalized Z Rotation 2nd Moment - Rotational Moment
16. Principal Axis Ratio

Table 2.1: Enumeration of Shape Descriptors calculated on Natural Ecotypes population

NASC Number	Accession Name	Longitude	Latitude	Country	Collector
N22603	CIBC-17	-0.6383	51.4083	UK	Mick Crawley
N22612	Uod-1	14.45	48.3	AUT	Marcus Koch
N22648	Ts-5	2.93056	41.7194	ESP	Albert Kranz
N22614	Cvi-0	-23.6167	15.1111	CPV	Albert Kranz
N22640	Mr-0	9.65	44.15	ITA	Albert Kranz
N22571	Pna-10	-86.3253	42.0945	USA	Joy Bergelson
N22582	Spr1-2	16	56.3	SWE	Magnus Nordborg
N22584	ÖMö2-1	15.7735	56.1509	SWE	Magnus Nordborg
N22631	Gy-0	2	49	FRA	Albert Kranz
N22642	Mt-0	22.46	32.34	LBY	Albert Kranz
N22649	Pro-0	-6	43.25	ESP	Joy Bergelson
N22616	Ei-2	6.3	50.3	GER	Albert Kranz
N22604	Tamm-2	23.5	60	FIN	Outi Savolainen
N22628	Br-0	16.6166	49.2	CZE	Albert Kranz
N22565	RRS-10	-86.4251	41.5609	USA	Joy Bergelson
N22591	Bor-4	16.2326	49.4013	CZE	Jirina Relichov
N22638	Kas-2	#N/A	#N/A	#N/A	#N/A
N22583	Spr1-6	14.1576	58.4173	SWE	Magnus Nordborg
N22624	Yo-0	-119.35	37.45	USA	Albert Kranz
N22645	Fei-0	-8.32	40.5	POR	Carlos Alonso-Blanco
N22587	Ull2-3	13.9707	56.0648	SWE	Magnus Nordborg
N22605	Tamm-27	23.5	60	FIN	Outi Savolainen
N22634	Ga-0	8	50.3	GER	Albert Kranz
N22581	Vår2-6	14334	55.58	SWE	Magnus Nordborg
N22564	RRS-7	-86.4251	41.5609	USA	Joy Bergelson
N22601	Sq-8	-0.6383	51.4083	UK	Mick Crawley
N22569	Rmx-A180	-86511	42036	USA	Joy Bergelson
N22577	Fäb-4	18.3174	63.0165	SWE	Magnus Nordborg

Table 2.2: List of Natural Accessions used in the GWAS experiment

. . . continued

NASC Number	Accession Name	Longitude	Latitude	Country	Collector
N22635	Mrk-0	9.3	49	GER	Albert Kranz
N22598	NFA-8	-0.6383	51.4083	UK	Mick Crawley
N22617	Gu-0	8	50.3	GER	Albert Kranz
N22589	Zdr-6	16.2544	49.3853	CZE	Jirina Relichov
N22578	Bil-5	18484	63324	SWE	Magnus Nordborg
N22611	Ren-11	-1.41	48.5	FRA	Gerhard Röbbelen
N22607	Kz-9	73.1	49.5	KAZ	Ihsan Al-Shehbaz
N22644	Wa-1	21	52.3	POL	Albert Kranz
N22590	Bor-1	16.2326	49.4013	CZE	Jirina Relichov
Edi1	Edi-0	-3.16028	55.9494	UK	Albert Kranz
N22595	Lp2-6	16.81	49.38	CZE	Ivo Cetl
N22659	Ws-2	30	52.3	RUS	Kenneth Feldmann
N22579	Bil-7	18484	63324	SWE	Magnus Nordborg
N22656	Bur-0	-6.2	54.1	IRL	D. Ratcliffe
N22597	HR-10	-0.6383	51.4083	UK	Mick Crawley
N22580	Vår2-1	14334	55.58	SWE	Magnus Nordborg
N22610	Ren-1	-1.41	48.5	FRA	Gerhard Röbbelen
N22574	Löv-1	18079	62801	SWE	Magnus Nordborg
N22570	Pna-17	-86.3253	42.0945	USA	Joy Bergelson
N22602	CIBC-5	-0.6383	51.4083	UK	Mick Crawley
N22627	Van-0	-123	49.3	CAN	Albert Kranz
N22643	Nok-3	4.45	52.24	NED	Albert Kranz
N22566	Kno-10	-86621	41.2816	USA	Joy Bergelson
N22567	Kno-18	-86621	41.2816	USA	Joy Bergelson
N22594	Lp2-2	16.81	49.38	CZE	Ivo Cetl
N22618	Ler-1	10.8719	47984	GER	Eric Holub
N22586	Ull2-5	13.9707	56.0648	SWE	Magnus Nordborg

Table 2.2: List of Natural Accessions used in the GWAS experiment

. . . continued

NASC Number	Accession Name	Longitude	Latitude	Country	Collector
N22626	An-1	4.4	51.2167	BEL	Albert Kranz
N22623	Ws-0	30	52.3	RUS	Albert Kranz
N22619	Nd-1	10	50	SUI	Eric Holub
N22625	Col-0	-92.3	38.3	USA	Albert Kranz
N22650	LL-0	2.49	41.59	ESP	Albert Kranz
N22637	Wt-5	9.3	52.3	GER	Albert Kranz
N22655	Ms-0	37.6322	55.7522	RUS	Albert Kranz
N22632	Ra-0	3.3	46	FRA	Albert Kranz
N22636	Mz-0	8.3	50.3	GER	Albert Kranz
N22653	Sorbo	68.48	38.35	TJK	Igor Vizir
N22641	Tsu-1	136.31	34.43	JPN	Eric Holub
N22613	Uod-7	14.45	48.3	AUT	Marcus Koch
N22568	Rmx-A02	-86511	42036	USA	Joy Bergelson
N22600	Sq-1	-0.6383	51.4083	UK	Mick Crawley
N22615	Lz-0	3.3	46	FRA	Albert Kranz
N22629	Est-1	25.3	58.3	RUS	Albert Kranz
N22622	Wei-0	8.26	47.25	SUI	Alan Slusarenko
N22647	Ts-1	2.93056	41.7194	ESP	Albert Kranz
N22606	Kz-1	73.1	49.5	KAZ	Ihsan Al-Shehbaz
N22609	Got-22	9.9355	51.5338	GER	Gerhard Röbbelen
N22652	Shahdara	68.48	38.35	TJK	Igor Vizir
N22573	Eden-2	18177	62877	SWE	Magnus Nordborg
N22575	Löv-5	18079	62801	SWE	Magnus Nordborg
N22658	Oy-0	6.13	60.23	NOR	Albert Kranz
N22593	Pu2-23	16.36	49.42	CZE	Ivo Cetl
N22620	C24	-8.42639	40.2077	POR	Brigitte Damm
N22592	Pu2-7	16.36	49.42	CZE	Ivo Cetl

Table 2.2: List of Natural Accessions used in the GWAS experiment

. . . continued

NASC Number	Accession Name	Longitude	Latitude	Country	Collector
N22621	CS22491	#N/A	#N/A	#N/A	#N/A
N22630	Ag-0	1.3	45	FRA	Albert Kranz
N22654	Kin-0	-85.37	44.46	USA	Albert Kranz
N22576	Fäb-2	18.3174	63.0165	SWE	Magnus Nordborg
N22596	HR-5	-0.6383	51.4083	UK	Mick Crawley
N22585	ÖMö2-3	15.7735	56.1509	SWE	Magnus Nordborg
N22588	Zdr-1	16.2544	49.3853	CZE	Jirina Relichov
N22633	Bay-0	11	49	GER	Albert Kranz
N22639	Ct-1	15	37.3	ITA	Albert Kranz
N22651	Kondara	68.49	38.48	TJK	Igor Vizir
N22572	Eden-1	18177	62877	SWE	Magnus Nordborg
N22599	NFA-10	-0.6383	51.4083	UK	Mick Crawley

Table 2.2: List 94 of Natural Accessions generated from Nordborg et al. (2002, 2005) and used in the GWAS experiment.

Area is a measurement of the "Rosette Projected Area', calculated in pixels but translated to squared millimetres. Circumference measures the perimeter of rosettes in pixels and it is also translated to millimetres. Similarly, Convex Hull Area and Circumference are measurements on the minimum surface surrounding the whole set of rosette pixels that not have any curves inward. Compactness is the ratio of Rosette Area over Convex Hull Area. Boundary point count is a pixel count of the perimeter, usually a sub-estimation due to using an 8-neighbourhood instead of 4-neighbourhood when choosing outer rosette pixels. Other measures of rosette extension are the radius of Minimum Enclosing Circle, the Area of the Minimum Rectangle and the maximum distance between rosette points, called Feret Diameter or Maximum Diameter or Caliper Length. Maximum Diameter correspond to the distance between the two farthest leaves tips.

As rosette morphology descriptors, several measurement of departure of a circle were calculated. Roundness measures the ratio between Area and squared perimeter. Roundness is also calculated from Boundary Point count and the convex hull roundness from Convex Hull Area and Convex Hull perimeter. Eccentricity is based in the statistical distribution of points, as well as rotational moment. The principal Axis Ratio, is the ratio between the two axis of an ellipse fitted to the rosette.

2.2.4 Image analysis

In order to calculate rosettes' shape descriptors automatically from images,they need to be processed using computer vision algorithms. Figure 2.3 illustrate the method used in this set of images.

Images were analysed at Lemnagrid software (Lemnatec, Germany) with a custom designed pipeline. The segmentation pipeline was split in three subroutines with complementary purposes, represented at figure 2.3). On one hand, greens pixels were segmented by calculating a grayscale image ($I_{Gray} = \frac{I^2_{Green}}{I_{Red}}$) from a Median Filtered image (Neighbourhood 11x11), and applying a threshold ($I_{Gray} > 65$). This method was good enough to separate rosettes from the background, however, pieces of bright paper, placed on pots border and barcode tags, remained in the image and the blue tray holder stay as well . In order to remove the bright tags, a different grayscale image is calculated ($I'_{Gray} = 0.33 \cdot I_{Red} + 0.67 \cdot I_{Green}$), now exaggerating the

brightness of pixels, and a threshold keeping only "not very bright" pixels ($I'_{Gray} > 133$). The resulting "Black and White" mask was *Dilated* with a square structuring element (size 10x10) for 5 iterations . A logical *AND* operation on the latter two mask provide a melted mask for rosettes.

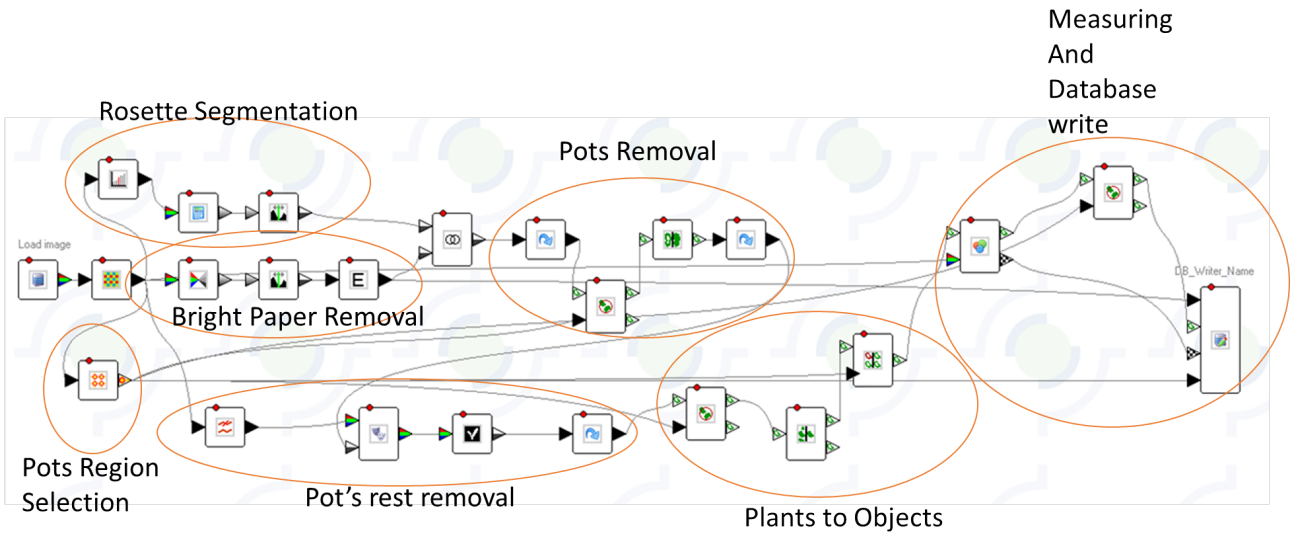
A third mask to filter out artefacts was made by actually masking the original image with the rosette-blue holder mask. Regions of Interest (ROI), i.e. hand-positioned circles to roughly determine plant position, were defined around each pot, so that blue holders remained out of those ROI. The rosettes plus artefacts were filtered with a lowpass filter,e.g. a Fourier transform-related filter that removes low frequency repetitions and blurs the image, (element size 5x5 and 5 iterations) and then applying a K-nearest neighbours to classify green pixels and remove black-brown ones. A final step was to filter out groups of pixels smaller than 100 pixels that corresponds to little soil and stones artefacts. All these steps result in rosettes silhouettes as final image.

Finally, the algorithm split the image in ROIs, as sub-images, and converted each rosette into "image objects", i.e a computer data structure that contains objects from an image and its properties. The object conversion cares about keeping leaves that overlap a neighbour pot into the original plant. Finally, for each rosette, shape descriptors are calculated and saved into Lemnatec database associated to tray and position.

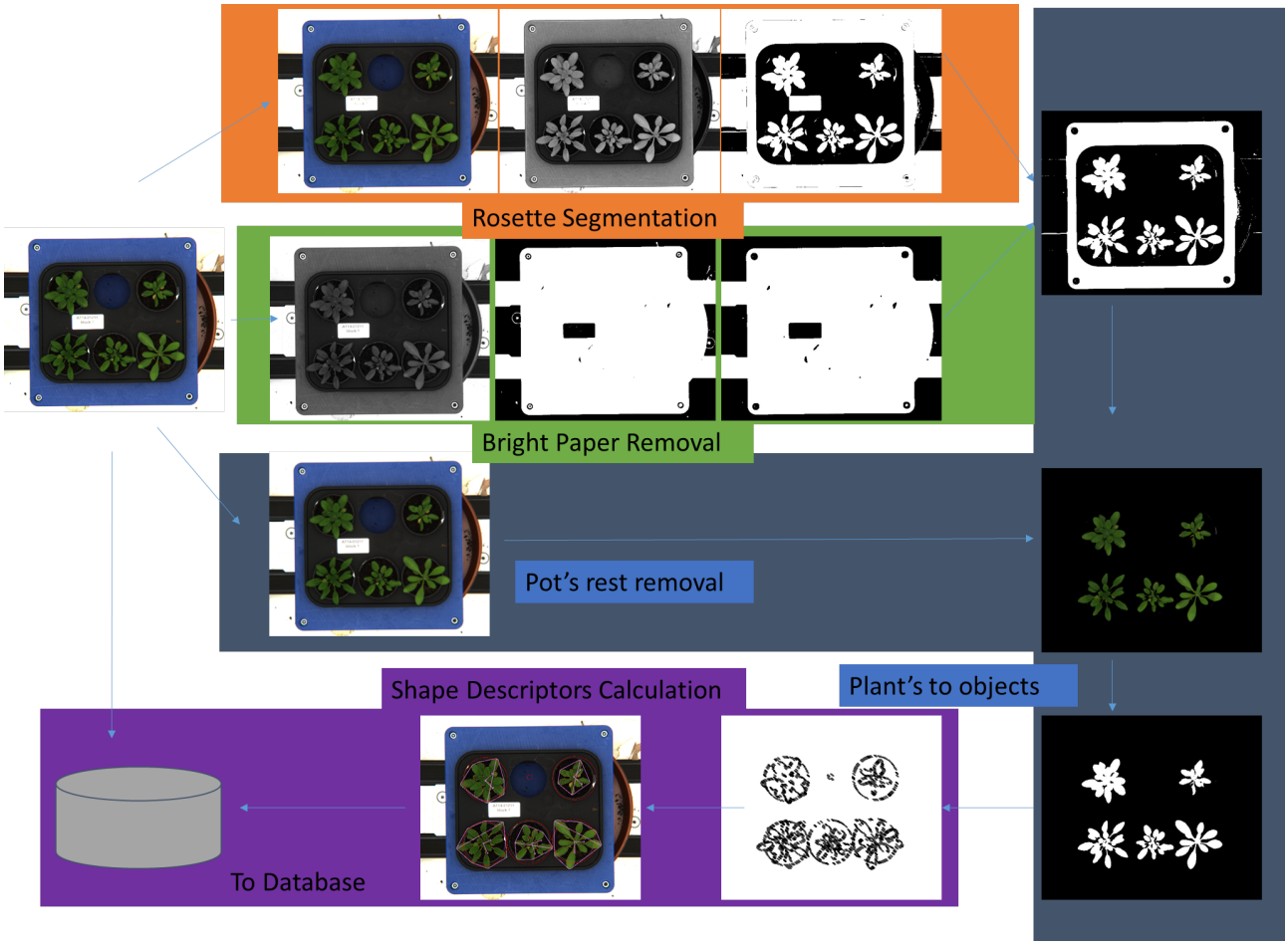
The described pipeline has, in general, a low False Positive Rate (artefacts classified as plant), but often some yellow-brownish leaves disappear, counting as False Negative Rate. Most of the descriptor are not strongly affected by this kind of small leaves missing.

2.2.5 Genome-wide Association mapping

GWAS mapping was performed using the software GAPIT as a package for R (Lipka et al., 2012). The function GAPIT implements several versions of the linear mixed model. Our choice was the "regular mixed linear model" instead of the more efficient "compressed mixed linear model" or "P3D/EMMAx" (Zhang et al., 2010). This selection was motivated to avoid a kinship-based clustering step in the compressed and EMMAx version (Tang et al., 2016; Yu et al., 2005) that seemed to artificially increase the estimation of heritabilities (data not shown). The model uses SNP alleles as fixed effect factors while correcting by population structure as



(a) Lemnagrid representation of image processing pipeline. Every box represent a function on the image. Boxes are connected representing a image processing workflow. The rosette is extracted from the background (Rosette segmentation) and in parallel, bright pixels from reflective tags are selected for removal. Bright paper is then removed from rosette segmentation and the resulting image is passed to a pipeline that remove pots borders that remain from the first step. Independently, the blue holder is classified in other pipeline and remove from the main resulting image. Finally, the rosette is the only object in the images and is passed through steps for conversion into a computer data structure needed for shape descriptor calculation.



(b) Example of segmentation in the Natural Accessions Experiment. Every tray is processed in three independent pipelines described in as described in a. First line shows the extraction of the rosette from their surroundings with some pieces of pot plastic and bright tags. Second line shows the extraction of bright tags and the intersection with the first line. The result is an image with rosettes, blue holder and nuisance pieces of pots. The third line represent the blue holder removal and the combination with the previous image to keep only the rosettes. Finally, rosettes are translated to a computer-friendly data structure that allows the calculation of geometrical properties needed to calculate the shape descriptors.

Figure 2.3: Example and description of Image segmentation. Panel a shows the workflow in the software Lemangrid and panel b represent an instance of a processed image

random effect. The algorithm, as we used, corrects the family structure by calculating a kinship matrix K, and population structure by a marker-based Principal Components Analysis matrix, Q (Patterson et al., 2006). The joint 'K+Q' approach is claimed to improve statistical power over the use of a single matrix approach (Lipka et al., 2012). For these analysis the latitude-longitude sampling coordinates for each ecotype was included as covariables in the model.

The package returns a complete analysis of phenotypes, markers and the statistical association between them. It allows to check Linkage Disequilibrium decay, PCA and Kinship Matrix as plots. Kinship matrix is represented as a heatmap with a dendogram for hierarchical clustering.In addition, the software plots phenotypic variability and its distribution for each phenotype. Association mapping is represented in Manhattan plots, that is, markers' physical position in chromosome against -log(p-value) (statistical significance of mixed model tests) and a Quantile-Quantile plot that allows to evaluate if the model accounts correctly for population structure (Korte and Farlow, 2013) and other covariates.

Numerical results include the determination coefficient, R^2, of every model with SNP and without it, corresponding to the full and null model, (Sun et al., 2010). P-values are reported raw and after False Discovery Rate correction is applied. The substraction of those 'determination coefficients' allow to calculate the percentage of explained variation by a SNP (Sun et al., 2010).

GAPIT calculates the Kinship Matrix using Van Raden's method (VanRaden, 2008). Principal Component Analisys is performed by the function prcomp in R by applying Single Value Decomposition(also called Q-mode PCA) (Zhao et al., 2007) on genotypic data. The PCA method uses natural accessions as observational units, and provide a linear combination on the markers' genotypes as explanatory variables. Markers are coded as homozygous (values = 0,2) or heterozygous(values =1), although ecotypes are considered as genome-wide homozygous). The result is a set of Principal Components (set up as 20 PCs) with values for each natural accession.

2.3 Results

Genome-wide Association studies results are strongly dependent on the chosen population and its structure. Initially, an overview of rosette shape descriptors results is provide. An analysis

of population parameters is provided before explore briefly some examples of GWAS results.

2.3.1 Phenotypic variation in size and shape descriptors

Our population shows visual and statistical phenotypic variation in rosette size and shape descriptors. Figures 2.6 and 2.7 shows four accessions, Ag-0, Col-0, Cvi-0 and Ler-1, with their replicates sort by ranking, according to Compactness values, (figures 2.4 and 2.5 and table 2.3). Figure 2.6 contains the individuals on the second day of phenotyping and figure 2.7 shows the same individuals, re-sorted by its ranking, on the 13th day of phenotyping. Rosettes from the same accessions are visually similar and different between accessions, so high heritability values are expected. Rankings by Compactness illustrates how the development extend the within accession variation For example, Ler-1 replicates are all very close at 2nd day, but get more sparse at 13th day. For Cvi-0, the development make them to pass from the 200th in the rank (average value) at 2nd day to 80th at 13th day, and less disperse so more homogeneous values.

Figure 2.8 shows Rosette Area, Rosette Compactness and Rosette Roundness time series for each plant, arranged by ecotype and with its mean value marked as a blue line and the standard error of the mean in pink. Size descriptors, like Area and Perimeter, follows the classical exponential or geometrical growth. However, shape-related descriptors time course is non-monotonic, i.e do not have a constant slope, that makes it difficult to model across time by simple curves. Due to that reason, a dynamic model approach is not followed, rather the decision of keep every trait and day as a single phenotype for GWAS was taken.

The correlation between descriptors (Figure 2.9) reveals an organization in three groups. A first group are size-related metrics accounting for Rosette area, maximum diameter (Feret's Diameter), and rosette coverage regions like convex hull area, minimum area rectangle area and minimum diameter circle. Also several metrics of rosette border are in this group, these are boundary points count, i.e raw count of number of pixels in rosette border, circumference, i.e distance covered by perimeter using the value of "1" for contiguous pixels and "$\sqrt{2}$" for diagonal pixels, and convex hull circumference. The second group include shape-related metrics accounting for rosette divergence from a circle. These descriptors are convex hull roundness (as $\frac{ConvexHullCircumference^2}{ConvexHullArea}$), boundary points roundness ($\frac{BoundaryPointscount^2}{Area}$), boundary points to area ratio ($\frac{BoundaryPointscount}{Area}$), eccentricity, the ratio of the two principal axis from second

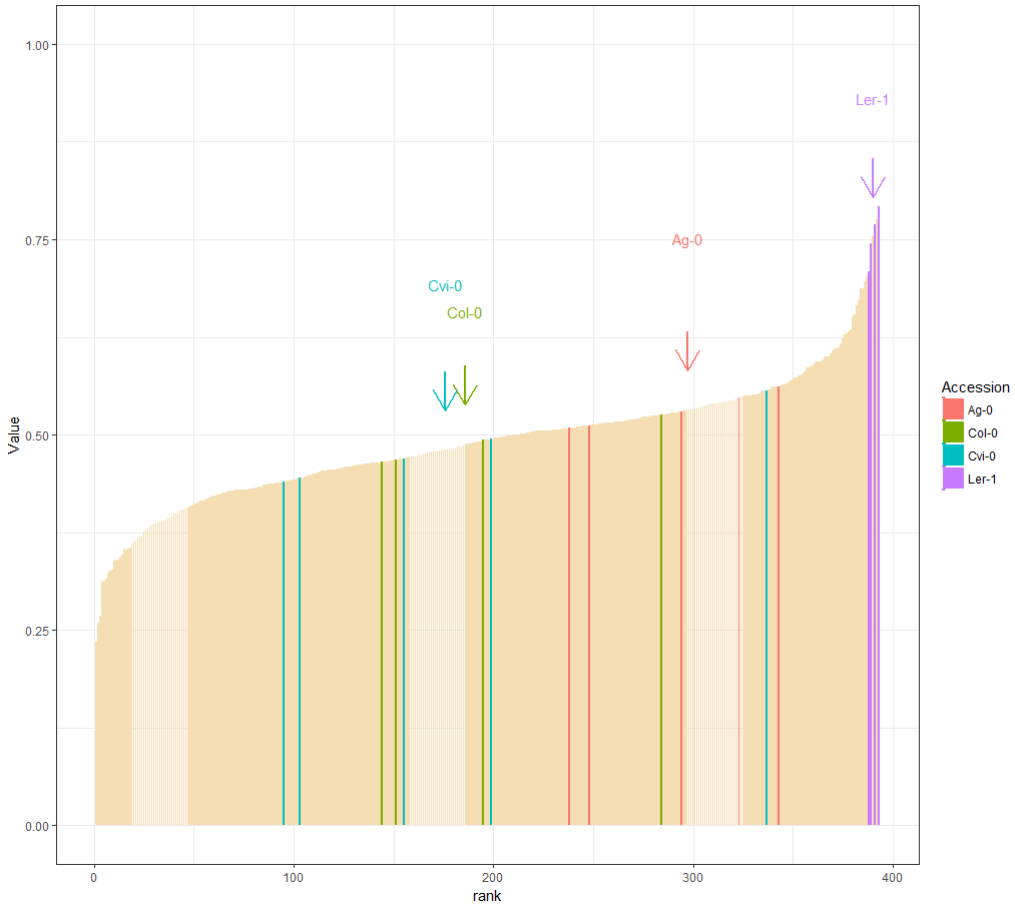


Figure 2.4: Ranking of Natural Accessions population by Compactness at DAE 2. Ranking of individuals. Colour represent accessions. Red = Ag-0;Green = Col-0;Blue=Cvi-0;Purple =Ler-0. The average value is indicated by an arrow and accession name

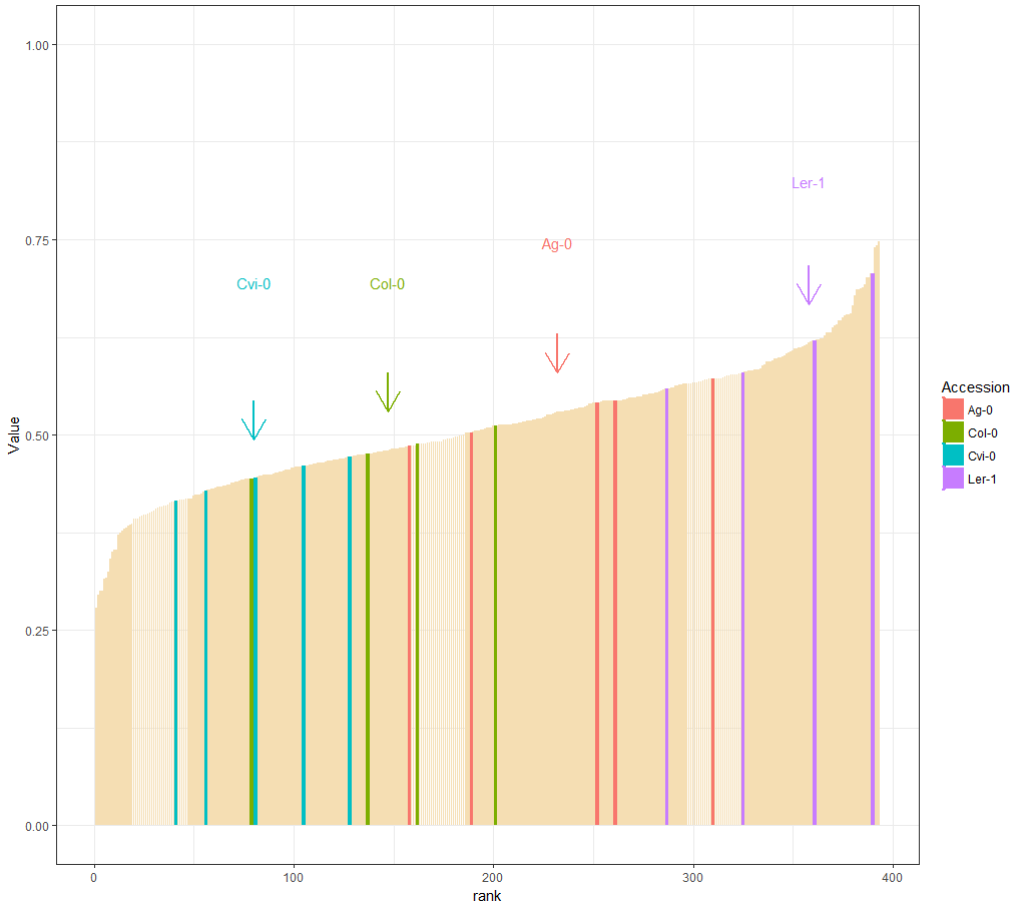


Figure 2.5: Ranking of the Natural Accessions population by Compactness at DAE 13. Ranking of individuals. Colour represent accessions. Red = Ag-0;Green = Col-0;Blue=Cvi-0;Purple =Ler-0. The average value is indicated by an arrow and accession name

Figure 2.6: Example of 4 Ecotypes, Ag-0, Col-0, Cvi-0 and Ler-1 from the Natural Accessions population, sorted by ranking at figure 2.4 and table 2.3. Each row contain 5 replicates (Ag-0 and Cvi-0) or 4 replicates (Col-0 and Ler-1) in the same order than table 2.3, sorted by Global rank. Thus. the more to the right, the more compact a rosette is in comparison with the rosettes in the same row. To compare between rows, the table 2.3 provides the required information.

Figure 2.7: Example of 4 Ecotypes, from Natural Accessions population, Ag-0, Col-0, Cvi-0 and Ler-1. Rosettes from same accession are in the same row, sorted by Global rank as shown at figure 2.5 and table 2.3. To compare between accessions, so between rows, the order is provided at table 2.3

Accession	Global Rank	Table Rank	Compactness	DAE	Barcode	Accession	Global Rank	Table Rank	Compactness	DAE	Barcode
Ag-0	238	8	0,51	2 days	AT14-74211	Ag-0	158	8	0,49	13 days	AT14-93211
Ag-0	248	9	0,51	2 days	AT14-93211	Ag-0	189	10	0,50	13 days	AT14-55211
Ag-0	294	11	0,53	2 days	AT14-36211	Ag-0	252	12	0,54	13 days	AT14-36211
Ag-0	323	12	0,55	2 days	AT14-17211	Ag-0	261	13	0,54	13 days	AT14-17211
Ag-0	343	14	0,56	2 days	AT14-55211	Ag-0	310	15	0,57	13 days	AT14-74211
Col-0	144	3	0,47	2 days	AT14-69211	Col-0	79	3	0,44	13 days	AT14-50211
Col-0	151	4	0,47	2 days	AT14-31211	Col-0	137	7	0,48	13 days	AT14-88211
Col-0	195	6	0,49	2 days	AT14-50211	Col-0	162	9	0,49	13 days	AT14-69211
Col-0	284	10	0,53	2 days	AT14-88211	Col-0	201	11	0,51	13 days	AT14-31211
Cvi-0	95	1	0,44	2 days	AT14-58211	Cvi-0	41	1	0,42	13 days	AT14-58211
Cvi-0	103	2	0,44	2 days	AT14-20211	Cvi-0	56	2	0,43	13 days	AT14-77211
Cvi-0	155	5	0,47	2 days	AT14-39211	Cvi-0	81	4	0,44	13 days	AT14-01211
Cvi-0	199	7	0,50	2 days	AT14-77211	Cvi-0	105	5	0,46	13 days	AT14-39211
Cvi-0	337	13	0,56	2 days	AT14-01211	Cvi-0	128	6	0,47	13 days	AT14-20211
Ler-1	388	15	0,71	2 days	AT14-68211	Ler-1	287	14	0,56	13 days	AT14-30211
Ler-1	389	16	0,74	2 days	AT14-30211	Ler-1	325	16	0,58	13 days	AT14-11211
Ler-1	391	17	0,77	2 days	AT14-87211	Ler-1	361	17	0,62	13 days	AT14-87211
Ler-1	393	18	0,79	2 days	AT14-11211	Ler-1	390	18	0,71	13 days	AT14-68211

Table 2.3: Sample of rosettes from DAE 2 and 13 from Accessions Ag-0, Col-0, Cvi-0 and Ler-1. Left side of the table contain the values of Compactess at DAE 2 with the ranking position at the whole population (Global rank), and in this table. Right side of the table contain similar information for DAE 13. Barcodes represent individual identifiers. There was 5 replicates of Ag-0 and Cvi-0 and 4 replicates of Col-0 and Ler-1

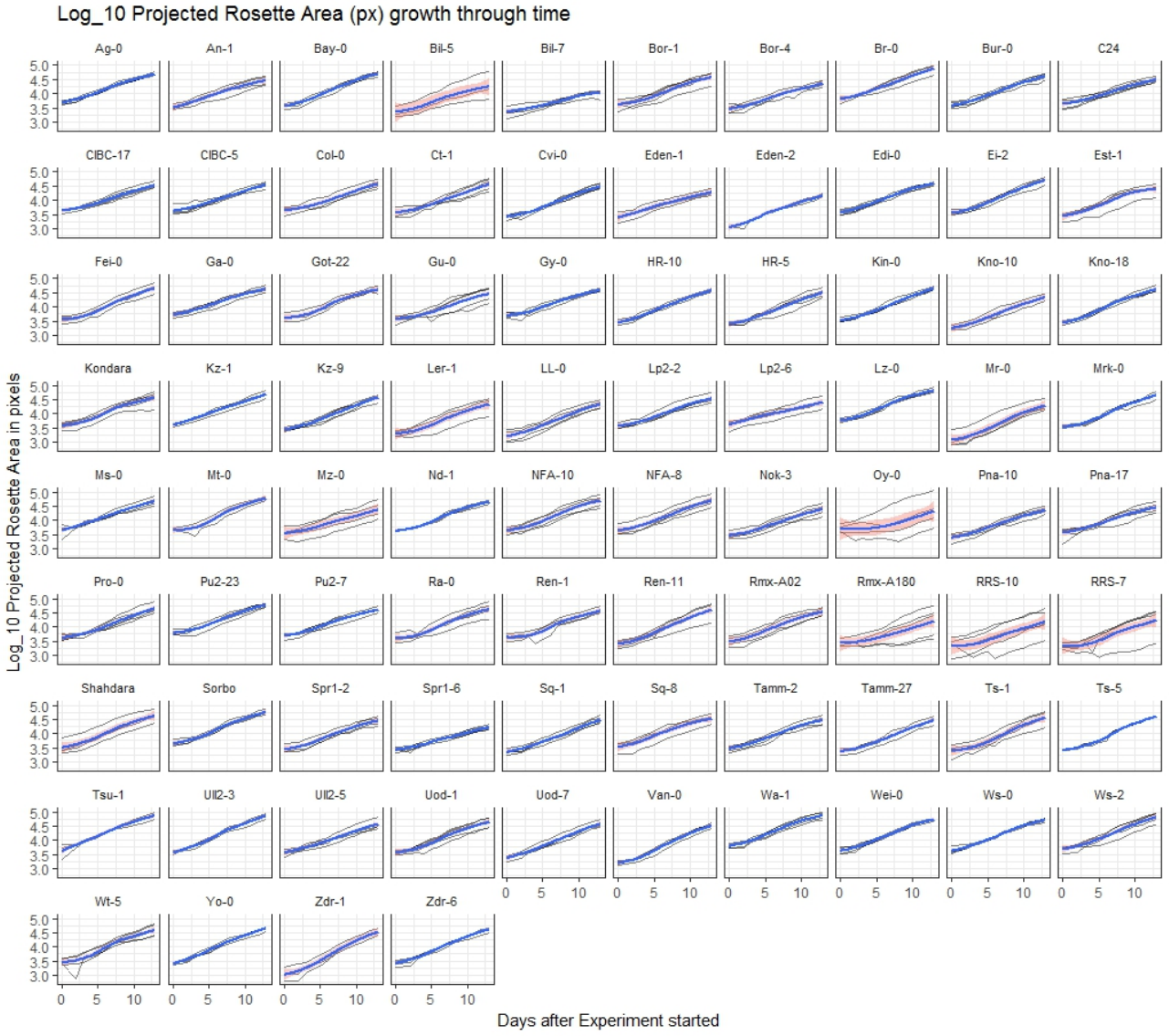
moments (the two latter are based in the point distribution of rosette pixels). The measurement of roundness, calculated as $\frac{Circumference^2}{Area}$ is also a departure from circle, shape-related, metrics, that present higher correlation with circumference than with other shape-related variables. Generally, each roundness measurement correlates with the primary parameters used for its calculation. Finally, a third group of measurements is compactness, $\frac{Area}{ConvexHullArea}$, that does not correlate with any other metric, and it accounts for how much the rosette is filling the region it covers. Interestingly, compactness does not correlates with their primary metrics, Area and Convex Hull Area, revealing that this ratio is not related with the size of rosette, neither influenced by it (as roundness does). Compactness is very influenced by the size and relative position of petioles, the closer the leaves, the more compact habit, and the longer petioles the more loose habit.

The different degree of correlation between similar traits, e.g roundness measurements, indicate that the different formulations captures rosette morphology in disparate ways. Therefore, they are not redundant, but add several perspectives to the same information source, forming all of them together a multivariate vector-valued shape metrics.

2.3.2 Kinship matrix

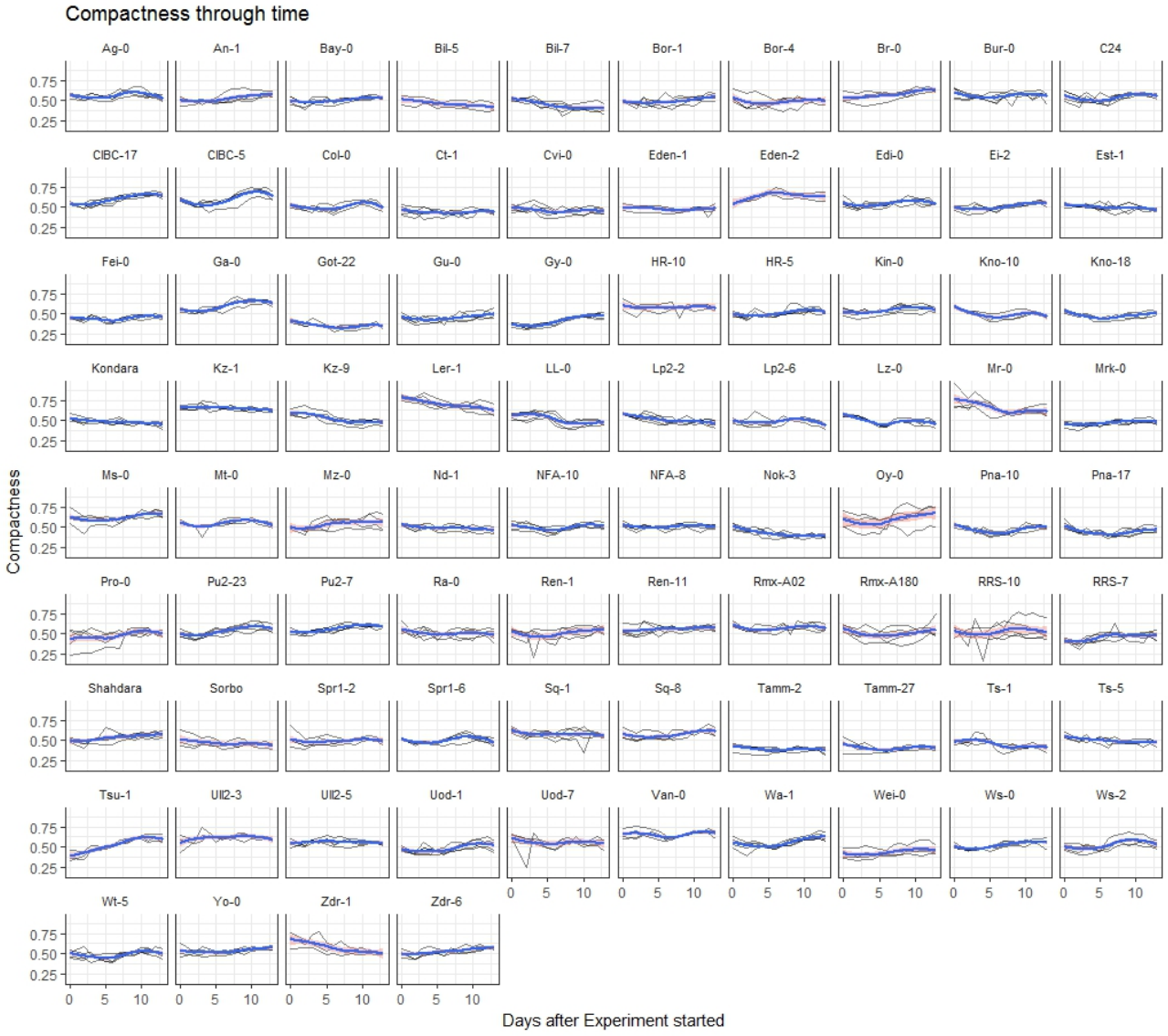
Four kinship matrices are shown at figure 2.10. Figure 2.10a represent the population structure for the whole Atwell's population, 199 ecotypes, at the full set of SNPs (21631). Figures 2.10b to 2.10d show the population structure of our experimental subpopulation of 91 ecotypes. It is calculated for the full set of SNPs, a subset of half of the markers equally spaced and a subset of a fifth of the markers equally spaces.

The cluster distribution of the total population, 199, individuals shows a separate cluster of the North American accessions, and two other identifiable clusters for the Central-European and Kazakhstan-Tajikistan populations. In the 91 accessions subpopulations, those clusters are even more apparent, either using the full set of SNPs or in any reduced version. These small clusters in the population can be explained by spatial proximity in their original sampling regions Specially those accessions that are subsamples of the same regions, e.g Zdr-1 and Zdr-2 coming from Zdar nad Zasavou in Czech Republic. Some clusters are nested into others, which is indicative of family and cryptic structure (Astle and Balding, 2009) that need to be corrected



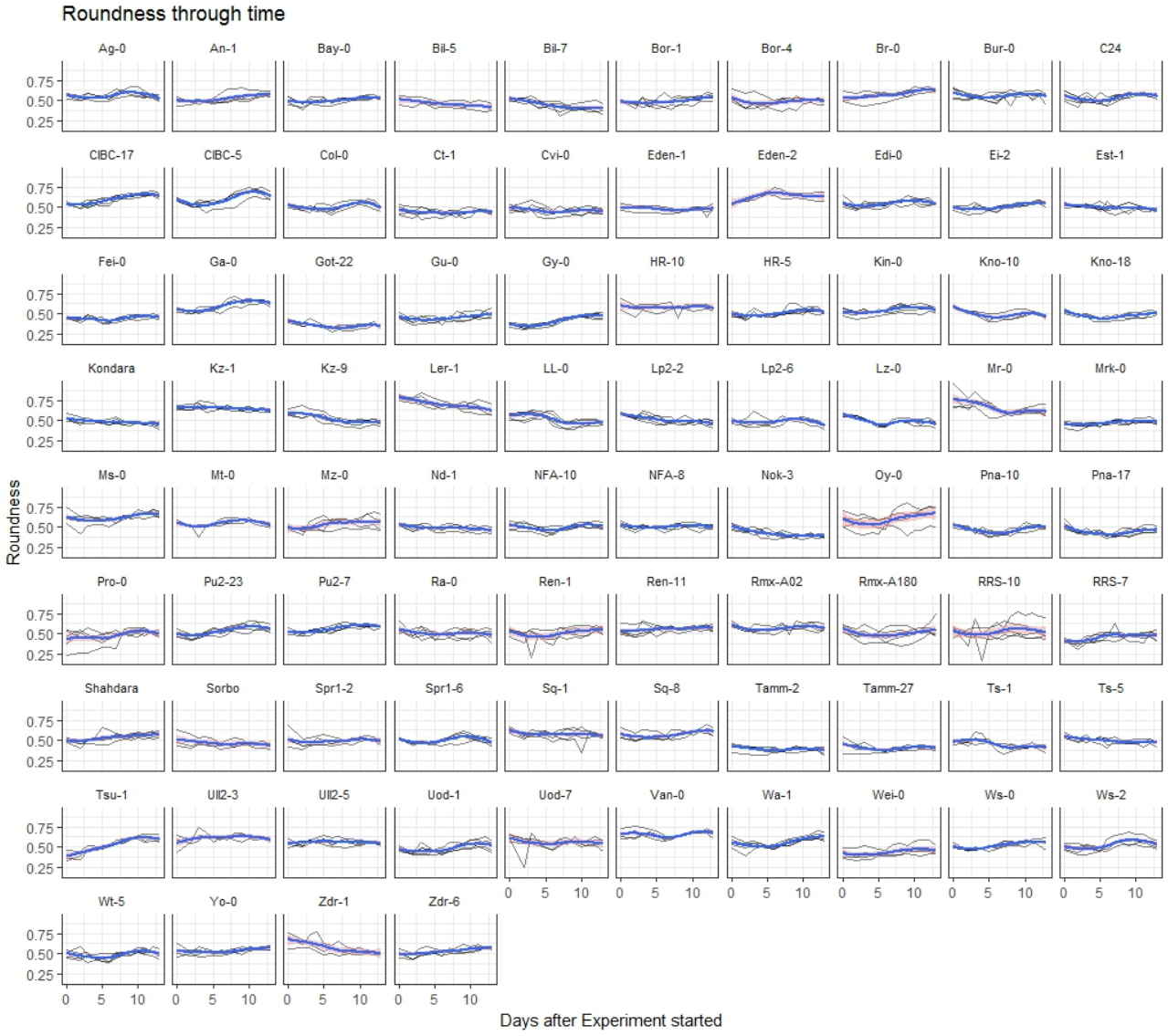
(a) Logarithm Rosette Area along time

Figure 2.8: Shape Descriptors across time by Accession
Loess smoothing in blue and variance in pink. It is equivalent to mean values and standard error of the mean.



(b) Rosette Compactness along time

Shape Descriptors across time by Accession
Loess smoothing in blue and variance in pink. It is equivalent to mean values and standard error of the mean.



(c) Rosette Roundness along time

Shape Descriptors across time by Accession
Loess smoothing in blue and variance in pink. It is equivalent to mean values and standard error of the mean.

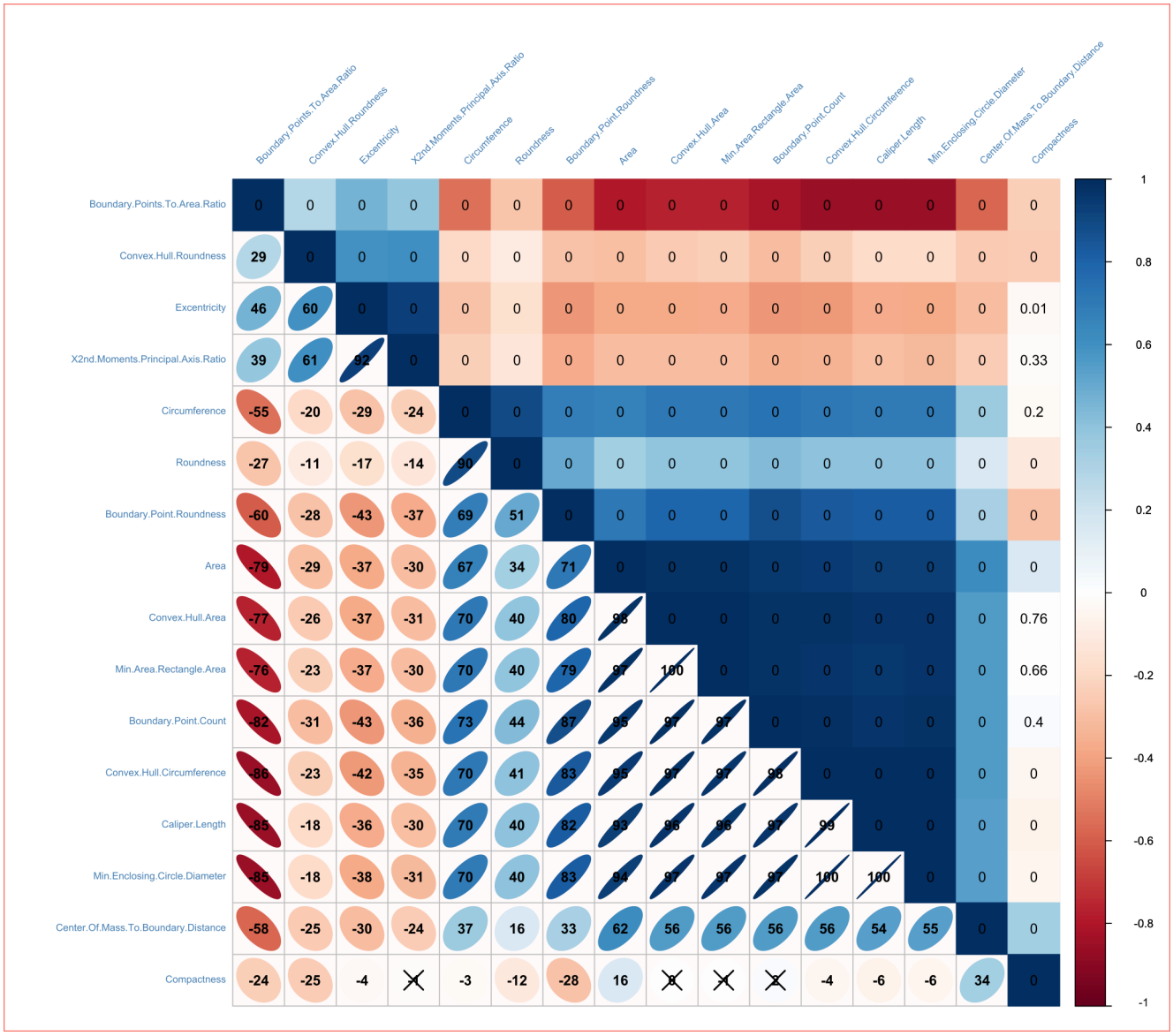
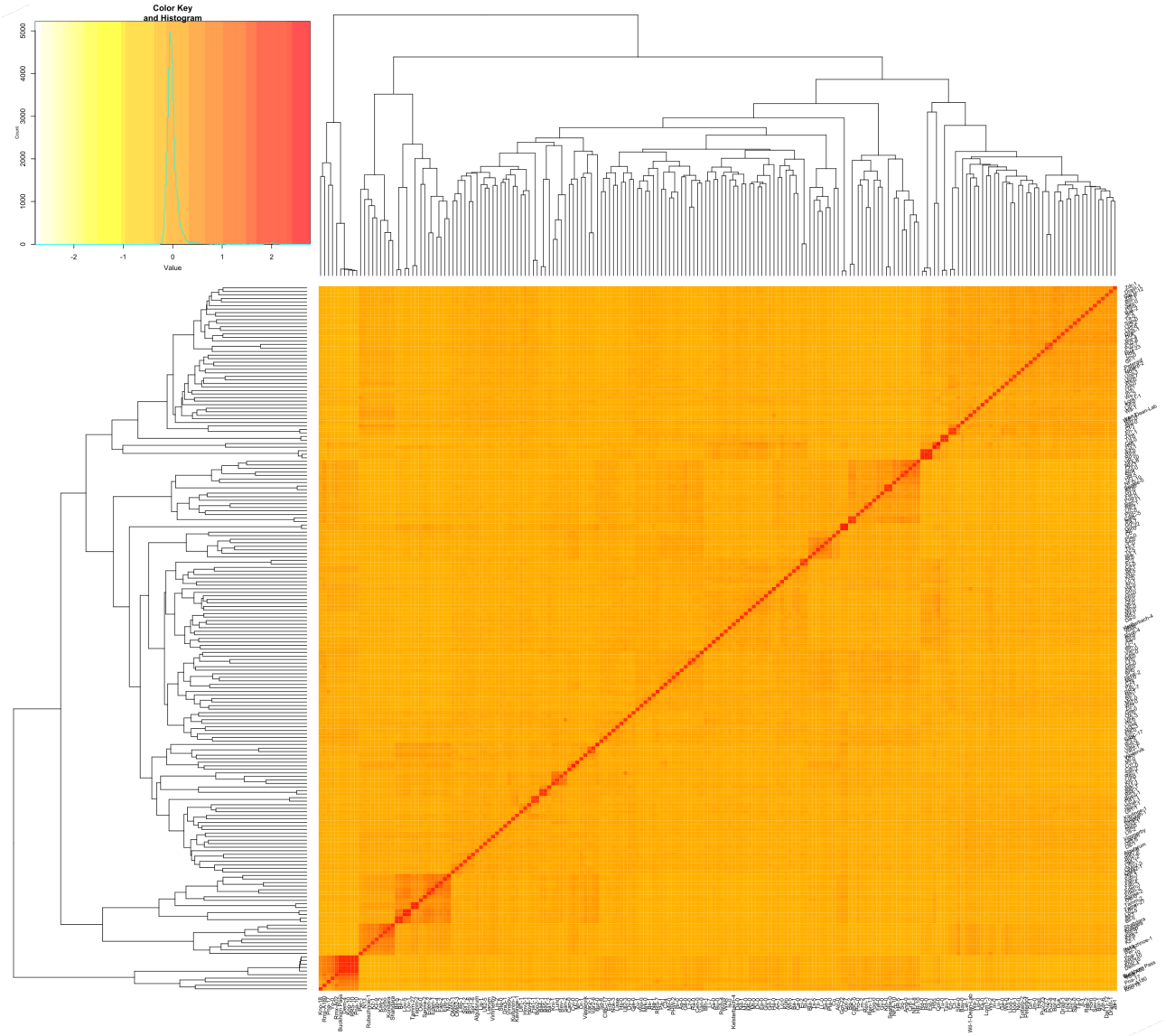
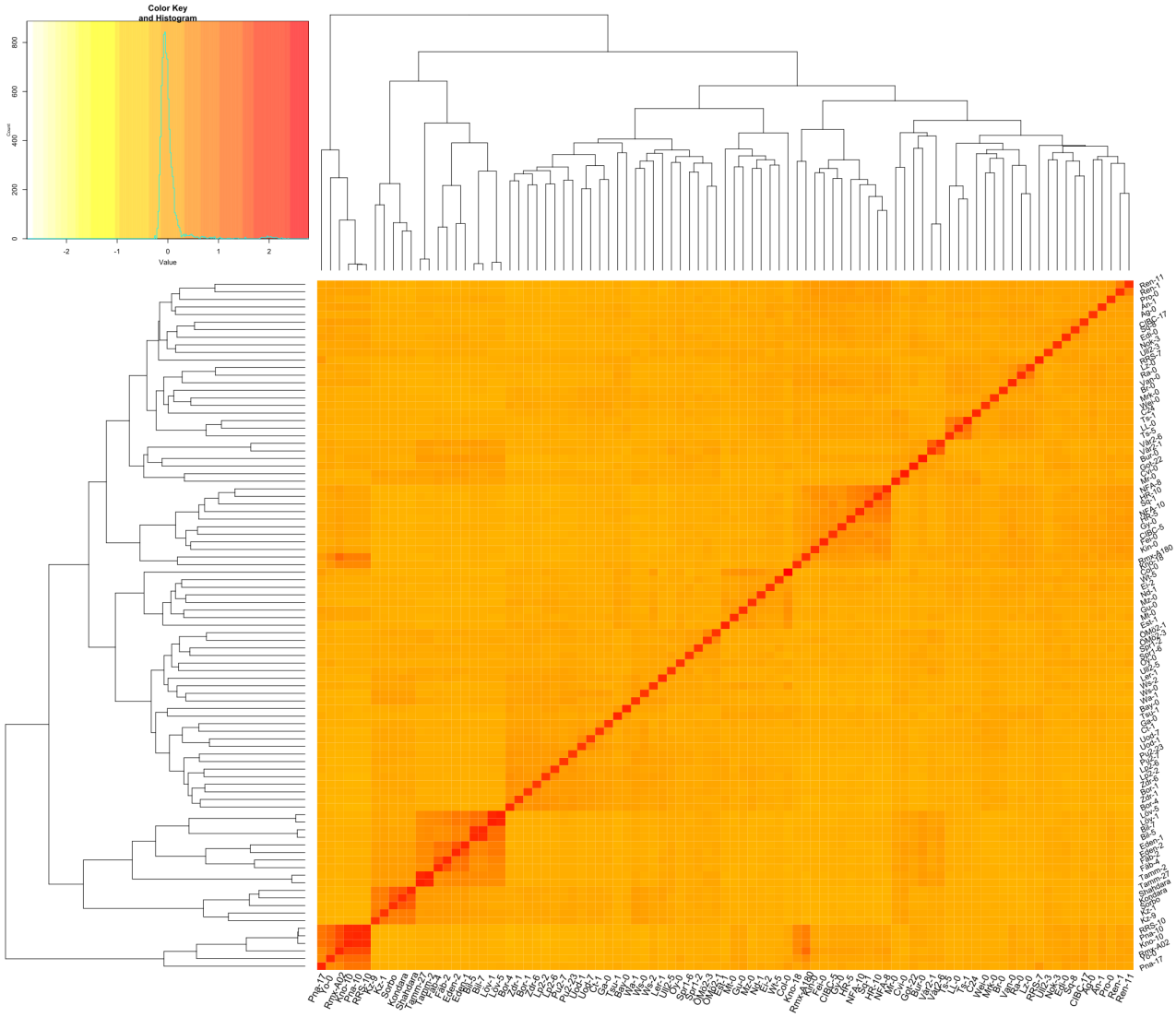


Figure 2.9: The four variables at top left are "Ratio Boundary points to Area", "Convex Hull Roundness", "Excentricity" and "Excentricity from ellipse 2nd moments". These four variables describes the similarity between the object and an ellipse, with higher values for ellipse-like objects, and smaller values for circle-like objects. This group correlate negatively with all other variables, that has smaller values when the rosettes are more ellipse-like. Compactness does not correlate with all the other variables.

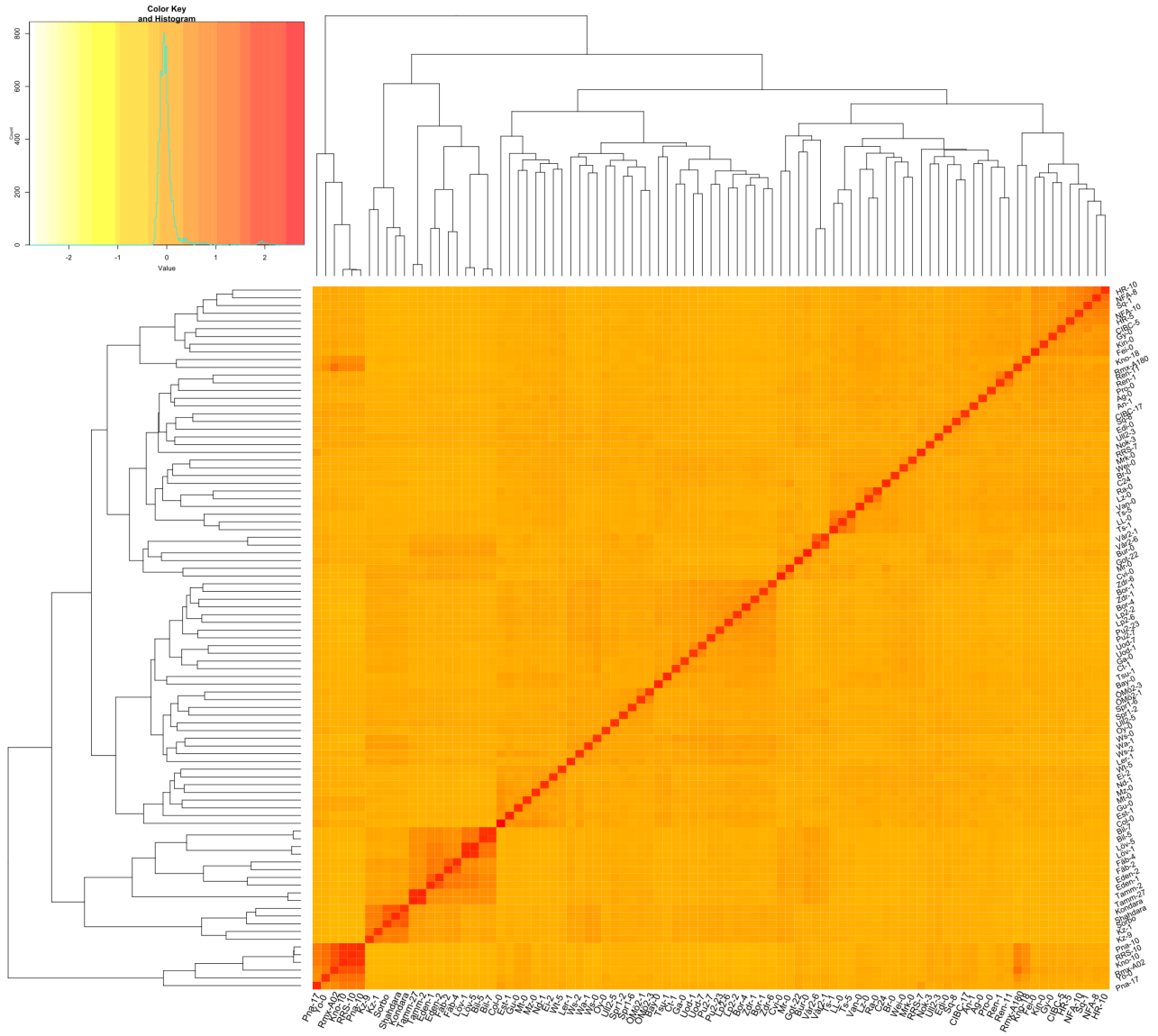
in the mixed model.



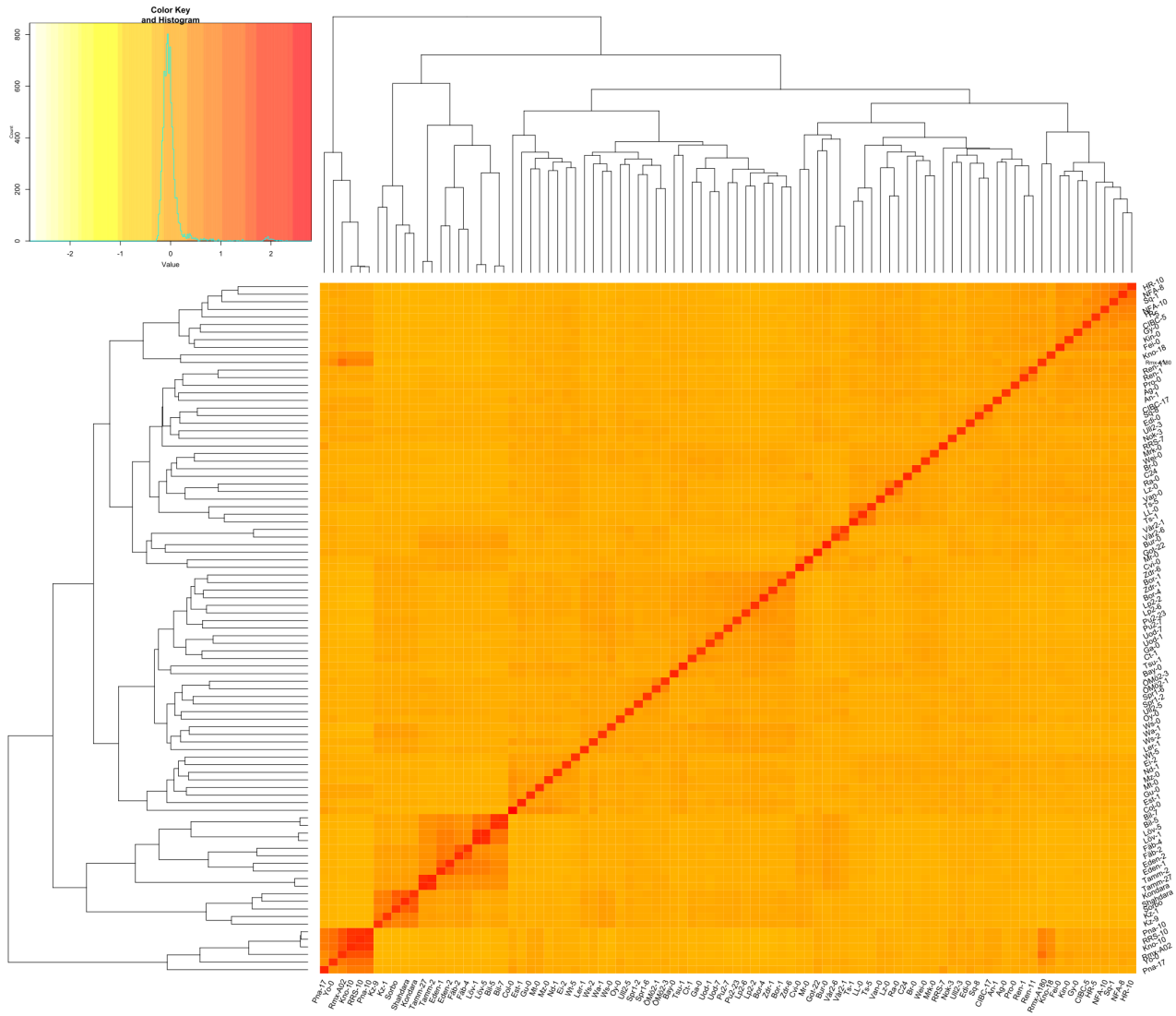
(a) Kinship Matrix of Natural Accessions population (199) from the full set of SNPs. The 199 accessions populations shows modules of accessions according to its origin with submodules. The cluster at bottom left is the American cluster, with accessions like Kno, Rmx or Buckhorn pass. Interestingly the subcluster (red rectangle) contains the same accessions that outside the rectangle, suggesting two overlapping populations. The second cluster at bottom left are accessions from Tajikistan, and the third one from Sweden. Other Nordic accessions are in the little cluster in the middle of the diagram, indicating population division. The clusters center-left of the picture correspond to groups of European accessions, being the bigger square from Germany, Poland and Czech Republic.



(b) Kinship Matrix of Natural Accessions subpopulation (91) from the full set of SNPs. The reduction of the population from 199 to 91 accessions maintains the clusters observed in the subfigure a, without representative change in the kinship matrix.



(c) Kinship Matrix of Natural Accessions subpopulation (91) from the half set of SNPs. The reduction to a 50% of the original amount of SNPs does not affect to the clusters corresponding to America, Tajikistan and Sweden, but reduce the kinship between European accessions.



(d) Kinship Matrix of Natural Accessions subpopulation (91) from the 5% set of SNPs. The reduction to the 5% of the original set of SNPs does not affect to the kinship matrix regarding figures a,b and c, so GWAS results would not be extremely affected by removing contiguous SNPs.

Figure 2.10: Four kinship matrix for 199 and 91 accessions using 216310 SNPs, 50% and 5% (continued)

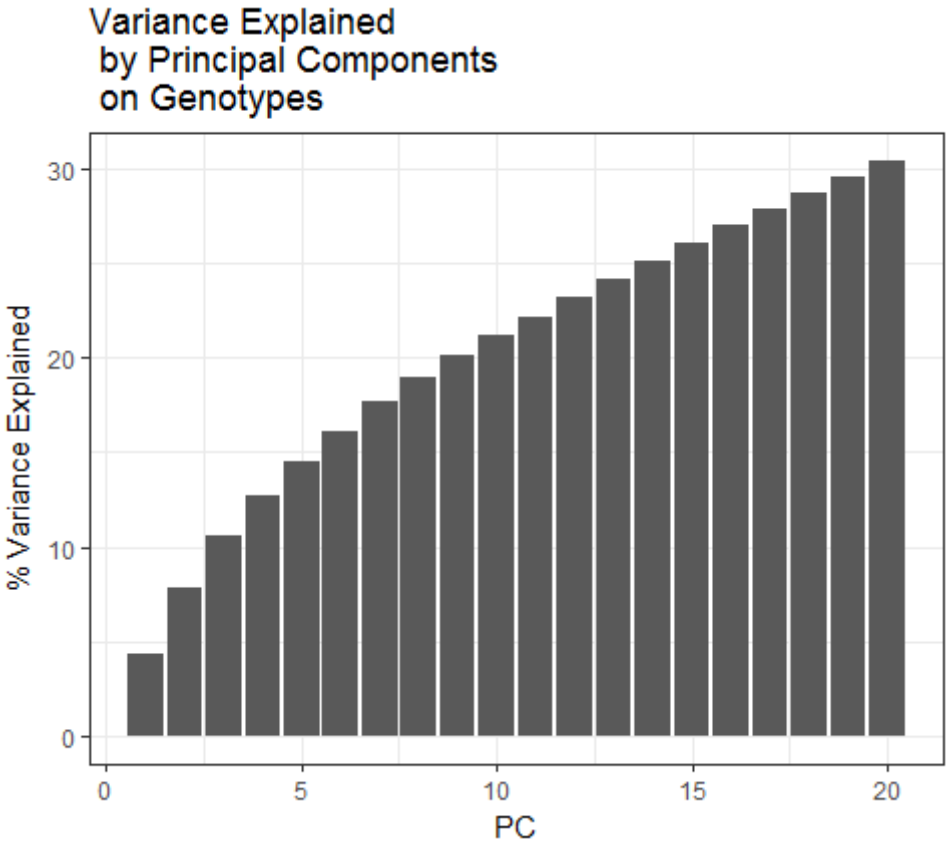
2.3.3 Principal Components on markers

The Principal Component method to account for population structure was able to explain up to $\sim$ 20% of variance for the 10 first PCs when using the whole set of markers in the complete population. The first PC only explains around $\sim$ 4% of the SNPs variation and the first 20 rise to $\sim$ 30% (figure 2.11a). When using 91 accessions and 5% of SNPs the first PC accounts for $\sim$ 6%, $\sim$ 30% for the first 10 PCs and rise to $\sim$ 42% for the first 20 PCs (figure 2.11b).

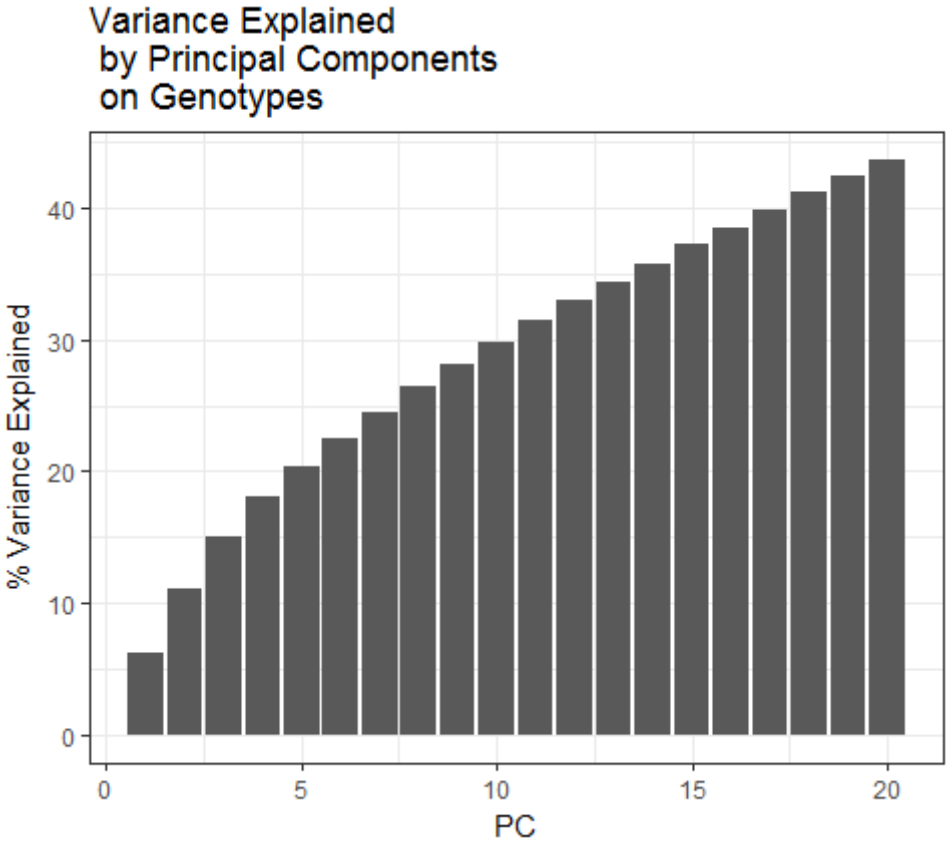
This results suggest that in spite of using almost half of the original population and using only a 5% of the SNPs available, the genetic diversity is almost similar and enough for GWAS mapping.

2.3.4 Linkage Dissequilibrium

GAPIT calculates LD between pairs of markers and return LD decay as a plot. Figure 2.12 illustrates the LD decay for the whole population (199 accessions, figure 2.12a) and our population of 91 accessions (figures 2.12b to 2.12d) using 100, 50 and 5% of the whole set of 216310 SNPs. Average LD using the whole and half set of SNPs incurs in the problem of a slower decay than previously published for this population, 10kb on average for the 199 accessions (Atwell et al., 2010) and 50 to 250 kb for the 91 accessions (Aranzana et al., 2005). Plots 2.13a and 2.13b contain the LD decay at the chromosome 5 for the 91 ecotypes population using the full set and the 5% set of SNPs. Plotting for each marker the distance to all markers that are in LD ($R^2 > 0.2$) provides a view of regions with larger LD and their extent. Thus, these plots illustrate the idea that, for the case of 91 natural accessions, the LD decay extends longer than 100kb in peaks concentrated in the centromeres and few other regions. When the number of SNPs is reduced to the 5%, the LD decay is closer to the 10kb values, but the high LD coldspots regions remains and have larger distance (between 2.5E05 and 1E06 bp). The extent of LD over 100Kb may strongly affect to the accuracy and ability to detect loci in this region associated to any phenotype, resulting in a increased false discovery rate.

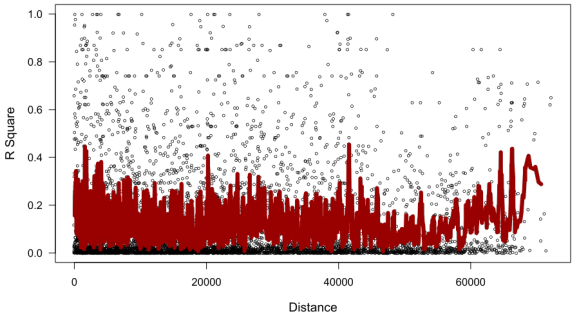


(a) 199 Ecotypes and full set of SNPs

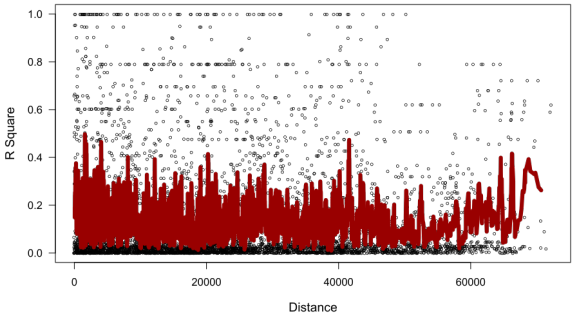


(b) 91 Ecotypes and 5%

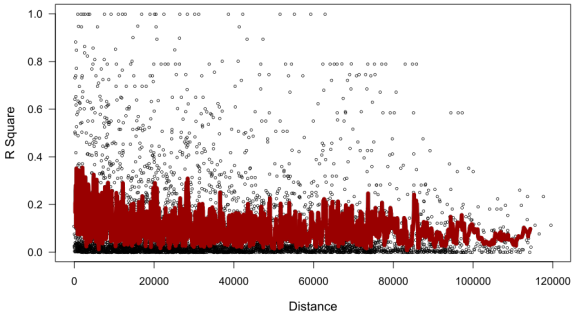
Figure 2.11: PCA of SNPs variation



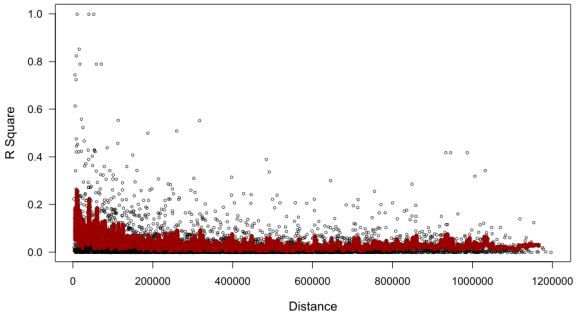
(a) 199 accessions - Full set of SNPs



(b) 91 accessions - Full set of SNPs

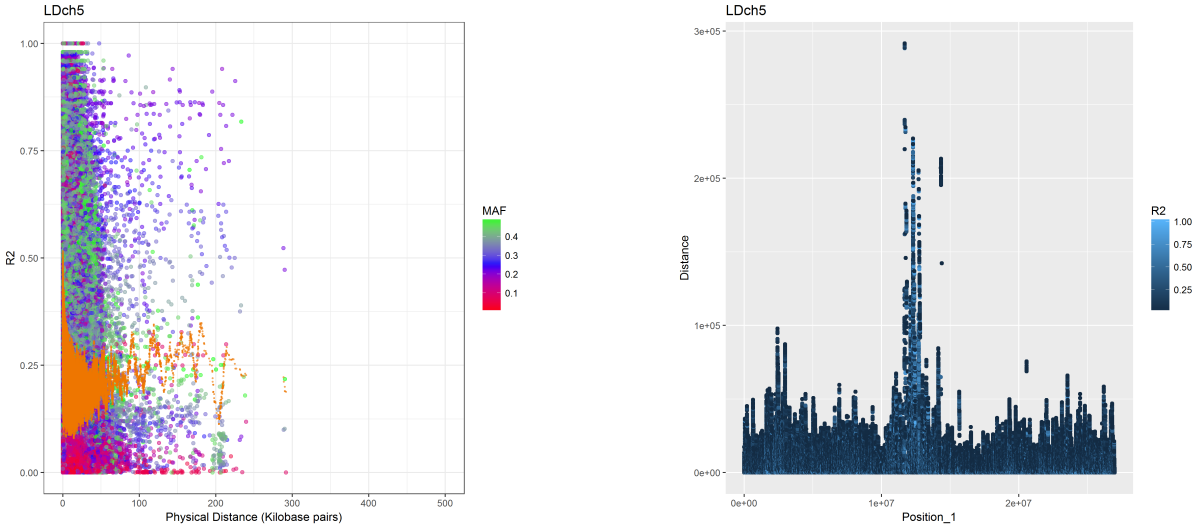


(c) 91 accessions - 50% set of SNPs



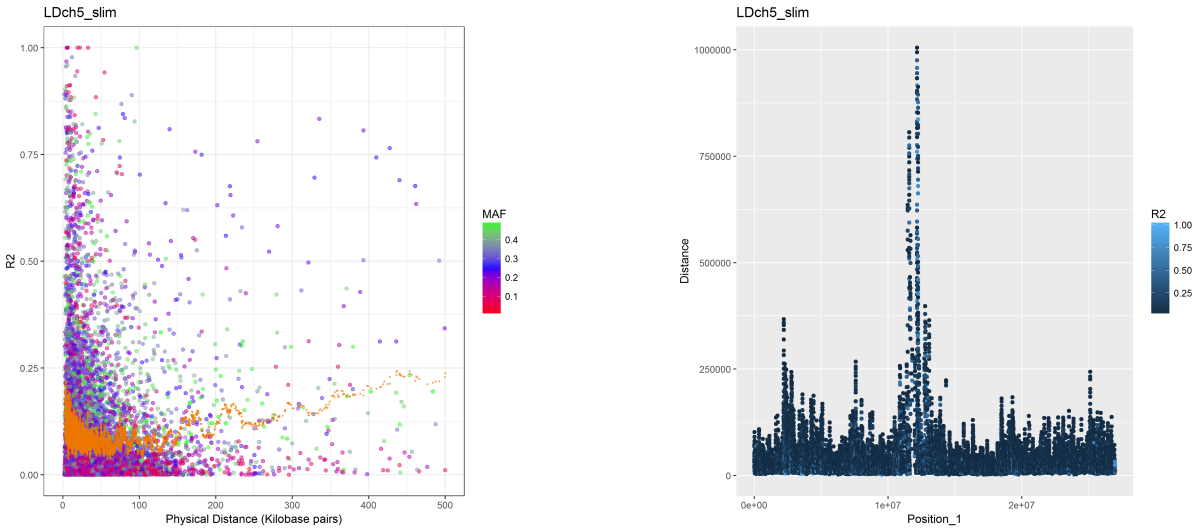
(d) 91 accessions - 5% set of SNPs

Figure 2.12: This figure represent the LD decay in the same four dataset that figure 2.10, showing that the excess in the number of SNPs affect the LD decay, having some SNPs a LD extent that covers long segments of the chromosome. Reducing the number of SNPs to teh 5% of the original number of SNPs reduce the LD extent, what would improve the resolution at QTL mapping.



(a) Whole set of SNPs.

The Linkage Disequilibrium Decay at the 91 Accessions sub-population, does not reflect an exponential decay pattern, yet many pairs of SNPs have high values of R^2 at distances as large as 200 kilobase pairs. Those long-range linked markers seems to be concentrated into regions, rather than homogeneously widespread along the chromosomes. In this example, the recombination cold spots seems to be around the centromere.



(b) Fifth set of SNPs - Chromosome 5

The Linkage Dissequilibrium Decay of a fifth of the whole set of SNPs at the 91 Accessions sub-population reflect a more conventional pattern. It Decays quickly in few Kilobase pairs, yet the recombination coldspots are still visible at the same regions that the previous plot. Using a fifth of the SNPs markers is expected to reduce spurious correlations due to long range LD.

Figure 2.13: Linkage Dissequilibrium Decay between markers in the 91 Accessions subpopulation - Example of Chromosome 5

2.3.5 Heritabilities

GAPIT estimate the genetic and environmental variance reporting them graphically. Broad sense heritability is calculated as $h^2 = \frac{V_g}{V_g+V_e}$. The procedure is repeated for every trait before GWAS is performed. Most size-related traits show high values of heritability (table 2.4), over 40% (yellow and red values) and generally decreasing through time. According to this, heritabilities for size measurements are generally high enough to consider that these phenotypes are repeatable within ecotype and that there is between ecotype variation.

Shape-related traits have, in general, lower heritabilities. Convex Hull Roundness, being a more robust version of robustness, had heritabilities of 0 most days, although rising to values up to 52%. Roundness, Convex Hull Roundness and Principal Axis Ratio are measurements of departure from a circular shape that generally were more noisy, explaining their low heritabilities. Some traits show high heritabilities of 100%. This values may be explained by the presence of extreme mean values, either quite high or quite low. It does not mean that they were outliers but simply values separate from the other accessions (for example the ecotype Var2-6 in figure 2.14).

2.3.6 Genome Association Mapping

For all the Shape Descriptors and DAE combinations, no marker showed any significant positive association with any phenotype. This was tested using the whole set of SNPs available, half of the set and a 5% of them. Models were calculated with and without using latitude and longitude geographical position as covariates . For the sake of brevity, not all those results are presented here. What follows is a general description of results from selected examples.

Trait_name	0	2	3	4	5	6	7	8	9	10	11	12	13
Area	97.34	96.50	100.00	69.42	91.31	84.74	83.56	77.50	78.47	80.33	80.28	80.05	81.98
Boundary.Point.Count	88.78	83.37	100.00	62.97	87.68	86.95	77.11	71.89	81.17	79.90	78.38	65.98	62.45
Boundary.Point.Roundness	51.70	67.88	85.17	63.96	88.14	100.00	66.34	84.36	89.46	93.54	94.47	84.77	80.30
Caliper.Length	91.33	99.45	100.00	50.22	72.52	63.96	57.53	58.42	60.86	59.10	66.49	55.77	68.12
Circumference	93.22	44.85	81.16	32.70	81.50	62.96	12.76	92.80	95.75	64.08	15.36	25.88	21.76
Compactness	78.50	100.00	80.07	33.29	71.02	61.43	44.99	44.05	57.48	98.15	99.71	100.00	85.55
Convex.Hull.Area	92.40	93.02	97.87	78.06	86.25	84.69	83.93	75.49	76.09	70.59	73.53	70.26	72.33
Convex.Hull.Circumference	85.74	92.00	93.72	68.41	75.54	71.78	68.19	61.61	61.36	64.16	71.67	66.24	64.67
Convex.Hull.Roundness	83.50	65.60	0.00	0.00	0.00	5.79	14.51	52.00	58.54	29.78	18.27	66.40	88.89
Excentricity	50.23	25.59	0.00	0.00	0.00	0.00	21.41	52.06	50.01	0.00	0.00	0.00	0.00
Min.Area.Rectangle.Area	95.79	92.59	98.28	77.34	80.33	84.46	87.63	80.31	71.08	68.51	76.56	71.28	78.58
Min.Enclosing.Circle.Diameter	90.84	97.92	100.00	49.66	71.09	65.87	61.38	60.10	59.55	62.67	69.92	64.86	65.29
Roundness	71.63	0.00	57.61	18.00	100.00	72.16	10.26	70.37	81.31	75.32	0.00	47.08	30.95
X2nd.Moments.Principal.Axis.Ratio	65.14	0.00	0.00	0.00	0.00	0.00	50.04	72.98	65.93	0.00	0.00	0.00	0.00

Table 2.4: The software GAPIT calculates heritability after a linear mixed model using the average value per accession, rather than the individual values. Extreme values, either 0 or 100, will occur for very small, or very large, residual variances in the model. Still the values are approximations to the heritability of the trait, providing the idea of consistency of such traits. Size-related traits, such Area and Caliper Length, has high values of h^2, while shape related has lower values, indicating higher variability between accessions with shared SNPs. Also, it can be observed that heritability of all traits decay along time for almost all traits. The shape-related trait Compactness has high values of h^2 for most of DAE and does not decay along time. This is indicative of being a traits suitable for finding shape-related QTLs through GWAS.

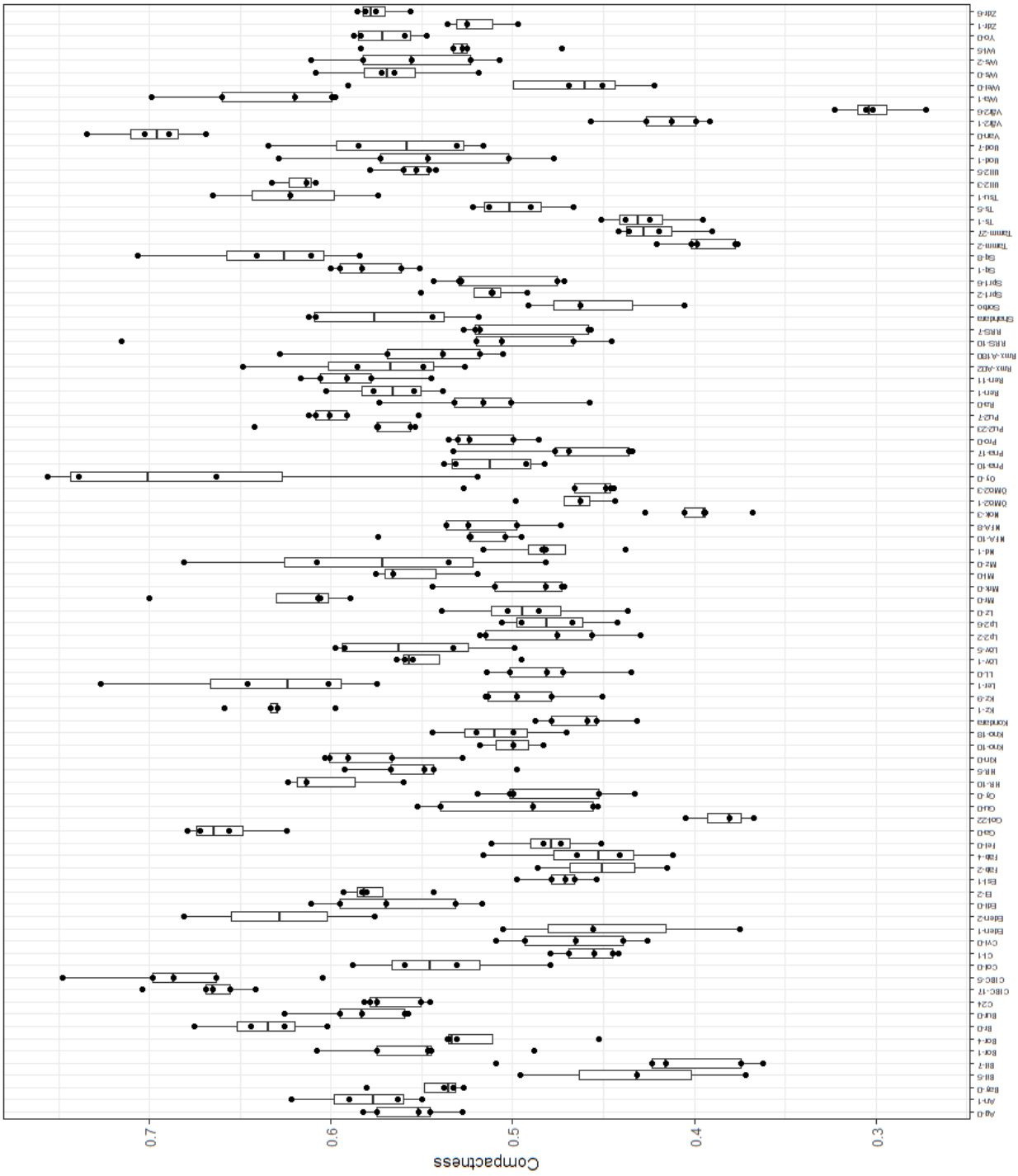
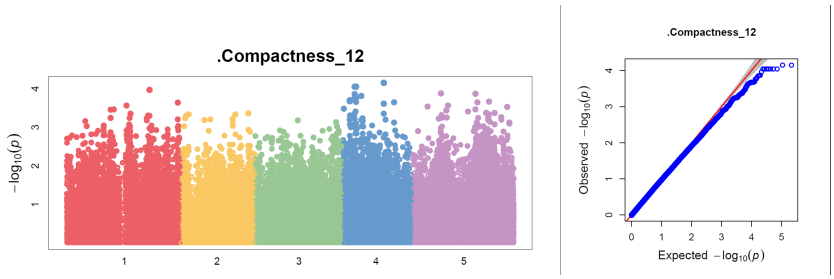


Figure 2.14: Boxplot and individual points for Compactness on DAE 12 by Accessions. Boxplots represents Median (horizontal bar), percentiles 25 and 75 (box edges). "Whiskers" are the limits from Tukey's formula for outliers ($P_{25} - (1.5 * (P_{75} - P_{25}))$ and $(P_{75} + (1.5 * (P_{75} - P_{25}))$. Circles are values beyond Tukey limits highlighted as potential outliers

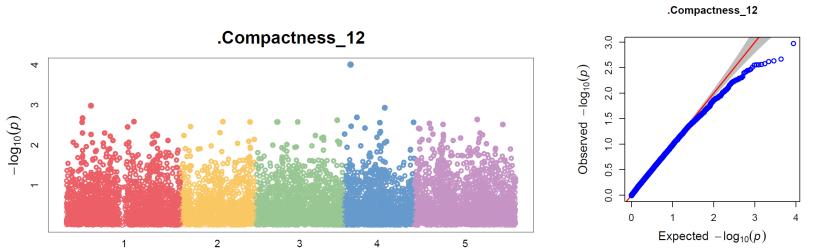
The trait Compactness 12 "Days After Experiment start" (DAE) illustrates the results of GWAS modelling of this dataset. This trait has a very high value for heritability of 100% although what follows is similar for traits with other heritability values. Figure 2.15 shows Genome-Wide Manhattan plots and P-values Quantile-Quantile plots for this trait using the whole (2.15a), and 5% of SNPs (figures 2.15b to 2.15d). For the model with the 5% of SNPs, three versions were computed. The first used the kinship matrix, the PCA-based population structure and geographical position as covariates. In the second, geographic location is removed from covariates but both kinship and population structure are kept. The third model keeps only the kinship matrix. The three models, for compactness_12 and for all the other traits, show no peaks neither over the False Discover Rate threshold nor Bonferroni correction threshold in the Manhattan plots. The highest p-value peak have FDR-p-values higher than 0.98, indicating they are far from being significantly positive. Looking to their corresponding p-values Quantile-Quantile plots, it is observed that all models are possibly over-corrected, since P-values are under their expected values. However, removing geography and population structure from covariates seemed to mitigate such overcorrection, although it is not enough. Correction for family structure was needed due to the results observed in the section "Kinship matrix", and was kept in all models. For PCA-based populations structure, first 20 PCs account no longer than 40% of genetic variation, suggesting it inclusion in the model should be required. Removing the Principal Component Analysis from the model correction slightly mitigate the effect but the model is still over corrected. It is possible to conclude that phenotypic variation was not distributed in such a way that any major or medium effect loci genetic variation captures its variability.

In spite of the lack of significance, some chromosome regions in the Manhattan plots looked as potential QTLs due to the distribution of p-values as peaks or "hills", and were consistent between days and even phenotypes.

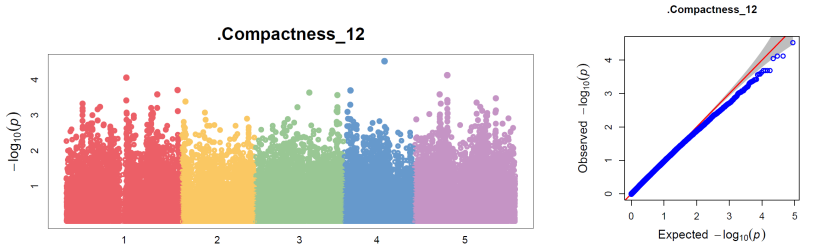
To further explore this regions a "joint" Manhattan plot for all phenotypes was built as follows. The p-values resulting from the model using only a 5% of SNPs with geographical origin, kinship and population structure correction ("K+P") were filtered over an arbitrary threshold of 7. When Manhattan plots from all phenotypes are plotted jointly, it becomes visible that certain SNPs are repeatedly at higher significance than others in their vicinity (Figures



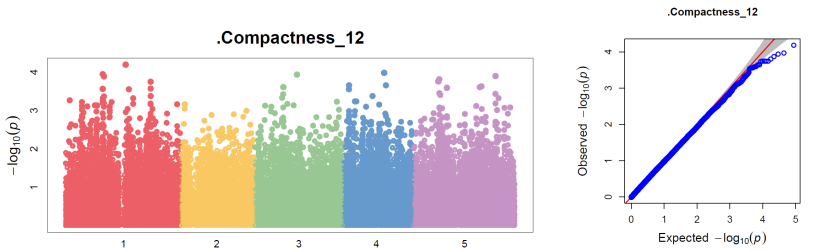
(a) Genome-Wide Manhattan plot - Full set of SNPs



(b) Genome-Wide Manhattan plot - 5% set of SNPs with Geographical position correction and PCA-based population structure correction



(c) Genome-Wide Manhattan plot - 5% set of SNPs without geographical position as covariables but with PCA-Based population structure correction



(d) Genome-Wide Manhattan plot - 5% set of SNPs and without geographical position as covariables nor PCA-based population structure correction

Figure 2.15: GWAS results for Compactness at 12 Days After the Experiment Start (DAE). Genome-Wide Manhattan Plots (left) and Quantile-Quantile plots (right) for a) Analysis using 216310 SNPs; b-d analysis with 10806 SNPS but different correction models. All models include kinship matrix as populations structure corrections.

2.16a and 2.16b). A histogram-like representation of the number of phenotypes (frequency) with -log(p-values) over 7 (figure 2.16b) indicates that some SNPs are hits for many phenotypes, for example a SNP at the end of the chromosome 1 has a -log(p-value) >7 for more than 40 phenotypes. A threshold on frequency of >10 phenotpes has been chosen to suggest potential QTLs for shape. That is, our suggested qShape are those SNPs with -log(p-value) over 7 in more than 10 phenotypes (figure 2.16b).

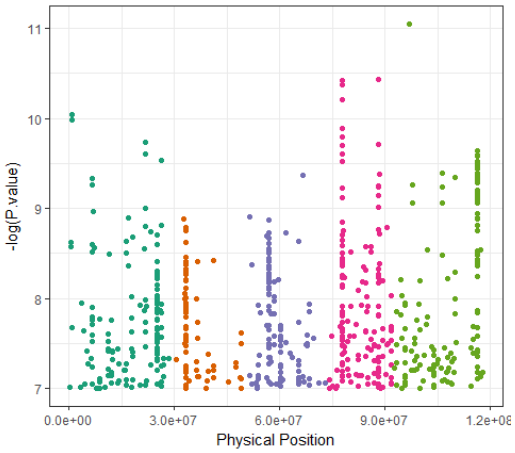
In summary, the fact that phenotypes showing diverse trends along time and between accessions pinpoint statistical association to allelic variation in the same SNPs suggest that, regardless the significance in this analysis, possible shape associated QTLs may exist in these positions.

According to this procedure 8 possible QTLs were identified (Table 2.5). To further study these regions, 8 intervals of 20kb (SNP position +− 10kb were formed. A search in TAIR9 annotated genome using *bedtools intersect* collect overlapping elements with such intervals (table 2.6). The selection only included the categories "gene","pseudogene" and "transposable element gene". In total 52 genetic elements were retrieved from the annotated genome. The AGI (Arabidopsis Genome Initiative) Identifier for genes is search in the TAIR9 bulk retrieval server [3] and araport thalemine server [4]. The search in TAIR ensure that the search is fetch in TAIR9 scaffold, but Thalemine provides clearer curated gene descriptions. Both servers provided similar results according to AGI indentifiers (table 2.7) quoting whether they are genes, pseudogenes or transposons and providing a brief descriptions that are printed in the table . A visual inspection of those gene activity did not bring any apparent functional element intuitively related with rosette growth or shape.

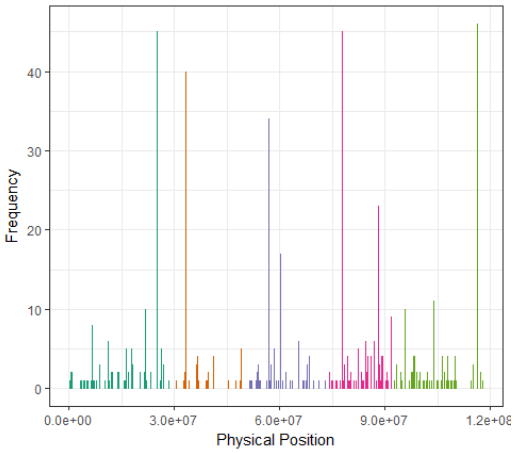
Chromosome	Position	QTL
1	24952802	Qtl1
2	2827219	Qtl2
3	6714023	Qtl3
3	9995074	Qtl4
4	4292939	Qtl5
4	14641845	Qtl6
5	11567666	Qtl7
5	24059497	Qtl8

Table 2.5: List of 8 possible QTLs

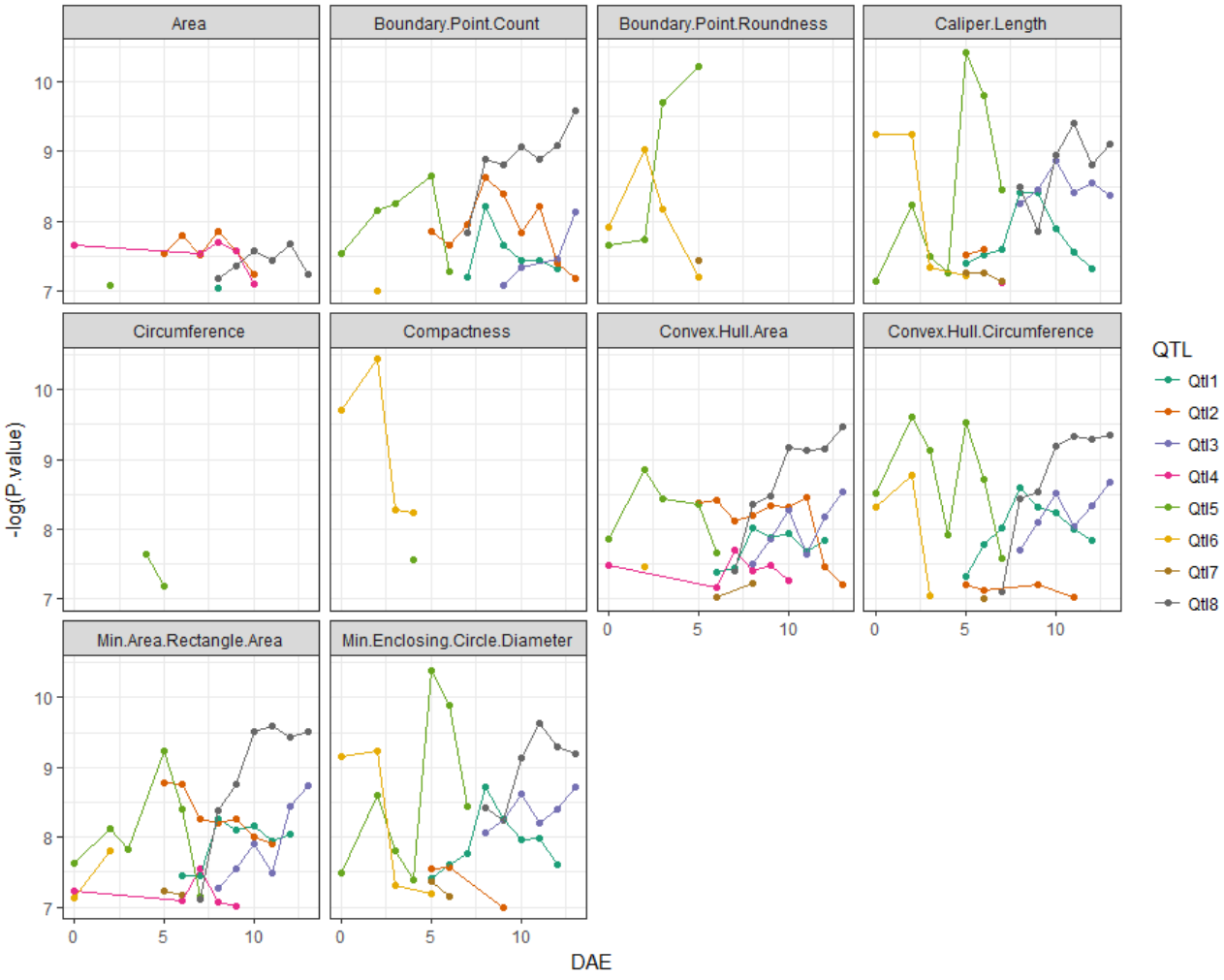
[3] `https://www.arabidopsis.org/tools/bulk/genes/index.jsp`
[4] `https://apps.araport.org/thalemine/begin.do`



(a) Joint Manhattan Plot for all phenotypes and DAE. SNPs were filtered out by a $-log(P-value) > 7$



(b) Count of the number of phenotypes_DAE where a SNP were over -log(P-value) over 7. 8 SNPs showed counts over 10, an arbitrary threshold to consider them as possible QTL



(c) Dynamic View of QTLs P-value for each Phenotype

Figure 2.16: Selection of QTLs from SNPs with high significance for several phenotypes and Days after Experiment started (DAE)

QTL	Chromosome	Initial Position	Final Position	AGI	Element Type
Qtl1	Chr1	24942187	24944497	AT1G66860	gene
Qtl1	Chr1	24945413	24945877	AT1G66870	gene
Qtl1	Chr1	24957557	24958114	AT1G66890	gene
Qtl1	Chr1	24959333	24961616	AT1G66900	gene
Qtl1	Chr1	24961633	24963946	AT1G66910	gene
Qtl1	Chr1	24946927	24955611	AT1G66880	gene
Qtl2	Chr2	2818438	2821477	AT2G06910	transposable_element_gene
Qtl2	Chr2	2826241	2826931	AT2G06912	pseudogene
Qtl2	Chr2	2827623	2828100	AT2G06914	transposable_element_gene
Qtl2	Chr2	2829210	2829798	AT2G06917	transposable_element_gene
Qtl2	Chr2	2837026	2838337	AT2G06922	transposable_element_gene
Qtl2	Chr2	2832082	2836514	AT2G06920	transposable_element_gene
Qtl3	Chr3	6701272	6704226	AT3G19340	gene
Qtl3	Chr3	6705492	6706127	AT3G19350	gene
Qtl3	Chr3	6707240	6709028	AT3G19360	gene
Qtl3	Chr3	6710719	6713650	AT3G19370	gene
Qtl3	Chr3	6714391	6716099	AT3G19380	gene
Qtl3	Chr3	6722988	6725027	AT3G19390	gene
Qtl4	Chr3	9985078	9986629	AT3G27080	gene
Qtl4	Chr3	9989539	9991826	AT3G27090	gene
Qtl4	Chr3	9991971	9994181	AT3G27095	pseudogene
Qtl4	Chr3	9994544	9996025	AT3G27100	gene
Qtl4	Chr3	9997894	10000230	AT3G27110	gene
Qtl4	Chr3	9999693	10003314	AT3G27120	gene
Qtl4	Chr3	10002658	10004281	AT3G27130	gene
Qtl5	Chr4	4283450	4283927	AT4G07493	transposable_element_gene
Qtl5	Chr4	4283942	4285470	AT4G07494	transposable_element_gene
Qtl5	Chr4	4288709	4291409	AT4G07495	transposable_element_gene
Qtl5	Chr4	4292652	4293196	AT4G07496	transposable_element_gene
Qtl5	Chr4	4296140	4297736	AT4G07498	transposable_element_gene
Qtl5	Chr4	4300048	4301484	AT4G07500	transposable_element_gene
Qtl5	Chr4	4301902	4305955	AT4G07502	transposable_element_gene
Qtl6	Chr4	14632652	14635885	AT4G29920	gene
Qtl6	Chr4	14648138	14653379	AT4G29940	gene

Table 2.6: List of overlaping Genes with QTls $\pm 20kb$

QTL	Chromosome	Initial Position	Final Position	AGI	Element Type
Qtl6	Chr4	14644008	14647591	AT4G29930	gene
Qtl7	Chr5	11555407	11558380	AT5G31511	transposable_element_gene
Qtl7	Chr5	11559202	11561951	AT5G30450	transposable_element_gene
Qtl7	Chr5	11563123	11564305	AT5G30721	transposable_element_gene
Qtl7	Chr5	11564676	11565897	AT5G30460	transposable_element_gene
Qtl7	Chr5	11566336	11566789	AT5G30648	transposable_element_gene
Qtl7	Chr5	11567515	11568304	AT5G30870	transposable_element_gene
Qtl7	Chr5	11569501	11572297	AT5G31092	transposable_element_gene
Qtl7	Chr5	11573251	11573608	AT5G31314	transposable_element_gene
Qtl7	Chr5	11575367	11575838	AT5G31536	transposable_element_gene
Qtl7	Chr5	11576744	11577128	AT5G31758	transposable_element_gene
Qtl8	Chr5	24046791	24050801	AT5G59680	gene
Qtl8	Chr5	24052384	24055425	AT5G59700	gene
Qtl8	Chr5	24057406	24061918	AT5G59710	gene
Qtl8	Chr5	24062576	24063277	AT5G59720	gene
Qtl8	Chr5	24063876	24066170	AT5G59730	gene
Qtl8	Chr5	24064477	24066154	AT5G59732	gene
Qtl8	Chr5	24051551	24052143	AT5G59690	gene

Table 2.6: List of overlaping Genes with QTls $\pm 20kb$ 20kb

AGI	Symbol	Description
AT1G66860	AT1G66860	Class I glutamine amidotransferase-like superfamily protein
AT1G66870	AT1G66870	Carbohydrate-binding X8 domain superfamily protein
AT1G66880	AT1G66880	Protein kinase superfamily protein
AT1G66890	AT1G66890	50S ribosomal-like protein
AT1G66900	AT1G66900	alpha/beta-Hydrolases superfamily protein
AT1G66910	AT1G66910	Protein kinase superfamily protein
AT2G06910	AT2G06910	transposable_element_gene
AT2G06912	AT2G06912	pseudogene of nucleic acid binding / zinc ion binding protein
AT2G06914	AT2G06914	transposable_element_gene
AT2G06917	AT2G06917	transposable_element_gene
AT2G06920	AT2G06920	transposable_element_gene
AT2G06922	AT2G06922	transposable_element_gene
AT3G19340	AT3G19340	aminopeptidase (DUF3754)
AT3G19350	MPC	maternally expressed pab C-terminal
AT3G19360	AT3G19360	Zinc finger (CCCH-type) family protein
AT3G19370	AT3G19370	filament-like protein (DUF869)
AT3G19380	PUB25	plant U-box 25
AT3G19390	AT3G19390	Granulin repeat cysteine protease family protein

Table 2.7: Araport Gene Description for the 8 qShape $\pm 20kb$

AGI	Symbol	Description
AT3G27080	TOM20-3	translocase of outer membrane 20 kDa subunit 3
AT3G27090	AT3G27090	DCD (Development and Cell Death) domain protein
AT3G27095	AT3G27095	pseudogene of Cysteine/Histidine-rich C1 domain family protein
AT3G27100	AT3G27100	transcription/mRNA export factor
AT3G27110	AT3G27110	Peptidase family M48 family protein
AT3G27120	AT3G27120	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT3G27130		
AT4G07493	AT4G07493	transposable_element_gene
AT4G07494	AT4G07494	transposable_element_gene
AT4G07495	AT4G07495	transposable_element_gene
AT4G07496	AT4G07496	transposable_element_gene
AT4G07498	AT4G07498	transposable_element_gene
AT4G07500	AT4G07500	transposable_element_gene
AT4G07502	AT4G07502	transposable_element_gene
AT4G29920	AT4G29920	Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein
AT4G29930	AT4G29930	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
AT4G29940	PRHA	pathogenesis related homeodomain protein A
AT5G30450	AT5G30450	transposable_element_gene

Table 2.7: Araport Gene Description for the 8 qShape $\pm 20kb$

AGI	Symbol	Description
AT5G30460	AT5G30460	transposable_element_gene
AT5G30648	AT5G30648	transposable_element_gene
AT5G30721	AT5G30721	transposable_element_gene
AT5G30870	AT5G30870	transposable_element_gene
AT5G31092	AT5G31092	transposable_element_gene
AT5G31314	AT5G31314	transposable_element_gene
AT5G31511	AT5G31511	transposable_element_gene
AT5G31536	AT5G31536	transposable_element_gene
AT5G31758	AT5G31758	transposable_element_gene
AT5G59680	AT5G59680	Leucine-rich repeat protein kinase family protein
AT5G59690	AT5G59690	Histone superfamily protein
AT5G59700	AT5G59700	Protein kinase superfamily protein
AT5G59710	VIP2	VIRE2 interacting protein 2
AT5G59720	HSP18.2	heat shock protein 18.2
AT5G59730	EXO70H7	exocyst subunit exo70 family protein H7
AT5G59732	AT5G59732	Natural antisense transcript overlaps with AT5G59730

Table 2.7: Araport Gene Description for the 8 qShape $\pm 20kb$

Finally, for the 8 possible QTLs, the p-value time dynamics has been explored graphically figure 2.16c. The aim of this is to check the relevance of time course at finding representative QTLs. It is observed that taking measurements through time increases the chances of finding QTLs. For example, QTL5, depicted in green, has its top p-value on DAE 5 for two disparate traits as Boundary Point Roundness and Minimum Enclosing Circle Diameter. Meanwhile QTL8, depicted in purple, is found after DAE 10, rising its p-values along time, for the traits related with region size as Convex Hull Area, Convex Hull Circumference, Minimum Area Rectangle Area and Minimum Enclosing Circle Diameter.

Traits Area and Circumference are not associated with high p-values, while the regions surrounding the rosette, i.e Convex Hull Area, Minimum Enclosing Circle Diameter and Minimum Rectangle Area, do provide more number of QTLs at higher significance. This happens in spite of Area and Circumferences are size-related and more robust to computer vision measurement artefacts. Therefore, this plot assist in the justification of using several geometrical measurements, although they are similar in meaning but different in the way it captures the shape of rosettes, as well as measuring them for several days and using each as separate phenotypes.

2.4 Discussion

For this study a subpopulation from the one used by Atwell et al. (2010) has been phenotyped for rosette morphological traits. Their publication represents a proof of concept of GWAS in Arabidopsis thaliana (Brachi et al., 2011; Korte and Farlow, 2013). This population was previously studied by Magnus Nordorg and collaborators concerning the degree of polymorphism, Linkage Dissequilibrium and its decay (Nordborg et al., 2002, 2005). Nordborg's population was used for association mapping, before to Atwell's paper, to find previously known genes, i.e candidate genes, for flowering time and pathogen resistance . The whole Atwell's population contains 199 accessions, and the LD decay was calculated in 10kb on average (see (Atwell et al., 2010; Korte and Farlow, 2013)). However, for the 96 ecotypes sub-population used here and by Nordborg et al. (2005), Aranzana et al. (2005) calculates a linkage dissequilibrium between 50 and 250kb.

For the 199 population, 140000 SNPs were calculated as sufficient for GWAS (Korte and Farlow, 2013), although the genotyping array used contain about 250000 (Atwell et al., 2010).

Atwell et al. (2010) studied 107 phenotypes with previously known candidate genes, testing the ability to find genes related with life history, e.g flowering time, pathogen resistance and ionomics. Atwell et al. (2010) mention that population structure plays a difficult role in this population and needs to be corrected. Similar findings in Nordborg's population were described by Aranzana et al. (2005); Zhao et al. (2007) when searching for known genes for flowering time. These authors argument that traits with geographical variation show higher number of spurious association, since plants from same origin may share more common variants and adaptations to their environment (Aranzana et al., 2005) . However, Vilhjálmsson and Nordborg (2012) pointed that the major reason of confounding at GWAS is the genetic background and environmental effects, over the population structure.

2.4.1 Population Structure

Our results about population structure agree with previous work (Aranzana et al., 2005) in which Linkage Dissequilbrium Decay extent further than 100kb, longer than in the bigger population studied in Atwell et al. (2010) for GWAS. However, Kim et al. (2007) studied LD in a population of 19 accessions and 341602 SNPs finding that recombination occurs often as hotspots in intergenic regions. Our results also suggest that LD is not homogeneously widespread along chromosomes, but located in "coldspots" around centromeres and certain regions. More analysis are needed in order to prove whether those longer linkage regions are either in intergenic regions or related with selection sweeps around certain genes (Kim et al., 2007).

In addition, Kim et al. (2007); Korte and Farlow (2013) indicated that 40% or 50% of the 25000 SNPs should be enough for GWAS studies in Arabidopsis. Our population structure calculations does not show visible differences whether using the full set of SNPs, 50% or 5% neither for the Kinship Matrix nor for PCA-based matrix, but the reduction down to about 10800 served to keep LD decay around 10kb in this population.

2.4.2 GWAS results

GWAS results shows that, for all the traits and "number of days after experiment started", no SNPs were found significantly associated with traits after FDR correction. Quantile-Quantile

plots showed that the EMMA models were over-corrected, in spite of whether PCA-based population structure correction and geographical origin were used as covariates or not. This result is unexpected since GWAS is know to have a high rate of false positive discovery, i.e spurious correlations, as a drawback(Hayes, 2013).

Possible explanations for this absence of significance is that morphological traits could have a genetic architecture composed of many genes of minor effects rather than rare variants with larger (Gibson, 2012) effects. However, it can be argued, for example, that the rare variant carried by the accession *Landsberg erecta* may show any significance in our data set, since it produces more compact rosettes and more round arrangement of leaves due to an strong allele in the gene *Erecta* (AT2G26330). The gene *Erecta* encodes for a receptor protein kinase that acts as modulator of environmental clues and regulates process related with plant physiology and development (Torii, 1996; Hall et al., 2007b; van Zanten et al., 2009b; Tisne et al., 2011; Shpak, 2013; Mandel et al., 2014).The results in this set of shape and size related traits may be explained by a conflated effect of common variants under the infinitesimal model, i.e the classical quantitative genetic assumption of additive effect of many genes (Barton et al., 2016, for a historical review), and the rare variants model, i.e major effect genes controlling a trait but found in few individuals(Gibson, 2012).

Similar results and conclusions were drawn by Kooke et al. (2016) from a GWAS on 349 accessions looking for the genetic architecture of morphological traits in Arabidopsis. They found extensive variation and high heritabilities on traits such as leaf length, petiole length, growth rate and inflorescence branching. They found few significant QTL, arguing that population structure may had lead to an increase of false negative. In addition, they interpret that part of the missing heritability in their analysis is due to small effect QTLs. In their words, the polygenic and highly complex genetic architecture of their traits "hides" the estimation of heritability. In their study, the presence of geographic and climatic adaptation explains phenotypic variation, making more difficult the elucidation of the genetic architecture. Finally, the authors mention that this effect should be even stronger for whole-plant phenotypes, such as the rosette morphology that I examined here.

Kooke et al. (2016) suggested that marker-based estimation and genomic prediction aids to identify genes when the trait is determined by small effect additive genes. Other approach

for complex traits is followed by Bac-Molenaar et al. (2015), modelling growth along time by exponential differential equation. This approach was not taken due to low differences among parameters for shape-related traits in previous pilot experiment (data not shown). This method would be more valuable for comparing growth rates among treatments as Bac-Molenaar et al did. However, a discussion about longitudinal data, i.e time series, is found at Li and Sillanpää (2015). In that review, several methods to account for time in "vector-valued" traits are proposed for increasing the opportunities of finding QTLs in GWAS by integrating the dynamic view from high-throughput phenomics data. In their results they explore graphically the temporal relation of traits and GWAS p-values.

In the same fashion than Li and Sillanpää (2015), my exploration shape descriptors as multivariate dynamic traits allowed to find a repetitive pattern over 8 SNPs. These QTLs lack significance, but remain having higher P-values for many traits and DAE than other SNPs regardless their vicinity. The visualization of p-values temporal course reflect that the genes underlying those QTLs become representative in a certain time frame.

It is suggested that the lack of significance might be due to the corrections on population structure and the number of SNPs. However, searching in the 20kb vicinity of those possible QTLs for genes, no well known genes with activities implying cell growth, signal transduction, hormone interaction or environmental response were found, e.g. phytochromes, flowering time genes, signal transduction factors, transcription factors, etc. Still, this 8 QTLs cannot be neither rejected nor accepted completely and further analysis on different populations should be required. Special attention over these 8 regions could help to elucidate the genetic architecture of rosette morphology as complex trait.

Further work could be done at checking the same traits on other population with different population genetic properties. In the QTL mapping literature biparental crosses of accessions grown in this experiment have been studies. Specifically, O'Neill et al. (2008) generated six biparental crosses, being one of them Ag-0 x Cvi-0. Those parental has shown different kind of rosettes in this experiment, so this cross is a candidate for QTL mapping of Arabidopsis rosette shape descriptors and it is explored in the next chapter.

Open questions, possibly of interest, are: a) to explain why some accessions with near origin and similar genetic composition, such as Var2-1 and Var2-6, has different values for

shape descriptors? b) Does latitudinal and longitudinal genetic variation correlates with rosette morphology?

Chapter 3

QTL mapping - Biparental Cross Cvi x Ag

3.1 Introduction

Recombinant Inbred Lines (RILs) from an cross of parentals Cvi-0 (Cape Verde Island) an Ag-0 (Argentat, France) Arabidopsis thaliana ecoytpes have been phenotyped for whole rosette morphological traits. Multiple QLT mapping (MQM) technique has been applied for gene mapping of Shape Descriptors.

The use of Biparental-derived RILs, being nearly genome-wide homozygous and phenotypically non-uniform population, is adequate for quantitative trait mapping (Weigel, 2011). A population derived from Cvi-0 and Ag-0 was chosen since parentals are relatively divergent lines for seedling rosette morphology, as observed in chapter 2 (see figures 2.4 and 2.5 and table 2.3).

The standard approach to QTL mapping using biparental crosses is testing whether a Quantitative Trait Locus is present and linked to a specific marker in the population under analysis (Haley and Knott, 1992). The methodology involve to identify a set of polymorphic genetic markers in parentals genome (Knapp et al., 1990; Knapp and Bridges, 1990), e.g Single Nucleotid Polymormisms (SNP), Simple Sequence Repeat (SSR) or insertion/deletions (InDels), that are homozygous for each parental but biallelic for both parentals.

Parentals are crossed to produce the first filial generation, noted as F1, which is heterozygous at all such loci. After subsequent selfing, crosses or backcrosses, noted as F_n, segregationresults

in a mixture of allelic combinations for all markers. Then, a genetic map is calculated from recombination frequencies between markers, being the distance the recombination probabilities expressed in centiMorgans (cM).

Afterwards, the association between a phenotype and genome-positioned markers is calculated to determine whether a marker is linked to a Quantitative trait locus for such phenotype or not (Broman, 2001) . Many techniques are available to calculate such association, sharing in common the use of markers as surrogate of their vicinity due to the linkage between markers and putative neighbour causal loci (Broman, 2001). The term linkage refers to a recombination frequency lower than 0.5 due to the closeness between markers and loci. Thus, the statistical association, calculated by linear model methods (Haley and Knott, 1992), t-test, Gaussian Mixtures (Lander and Botstein, 1989) or any other method (Knapp et al., 1990), between a marker and a phenotype indicates whether a possible QTL is nearby the marker.

In general, a normally distributed phenotype in a population is a mixture of several normal distributions according to allelic variation at any causal QTL. Maximum Likelihood methods separate the empirical phenotypic distribution into Gaussian distributions according to allelic values at a single marker (Lander and Botstein, 1989). Likelihood Ratio (LR) test quantifies the "effect size" between allelic phenotypic means, resulting in a LOD score (Log of Odds = LOD = -log10(p-value)) . Interval mapping is a refinement over this model that improves precision by locating a QTL within intervals between flanking markers (Jansen and Stam, 1994). It involves to calculate the allelic frequencies of a potential QTL between two markers according to the local recombination frequency and compute the mixture distribution between estimated genotypes at QTL being tested (Zou and Zeng, 2008). This approach uses pseudo-markers, positions in the chromosome that has not been genotyped, but whose genotype is calculated from flanking markers genotypes. In spite of changing the computation, LOD scores keep the usual interpretation.

The previously explained approach only allows for testing a marker at a time. Instead, using some markers as covariates and analysing simultaneously sets of markers increase the precision and reduce false positives. Multiple QTL mapping (MQM) (Jansen and Stam, 1994) uses multiple linear regression methods to regress the phenotype over markers and putative QTLs. The MQM, as implemented in the package R/QTL, proceed as follows (Broman

and Sen (2009) for a more detailed explanation). Population genotypes need to be first "augmented", meaning that missing genotypes are predicted by imputation. Then, a set of markers are selected as cofactors, either automatically or chosen by the user . The automatic selection takes into account marker density for a equally disperse sample. These cofactors are to be used as covariables in the linear regression. Finally, MQM proceeds to scan groups of markers for phenotype-QTL association, accounting for variation in the cofactors and removing them "backwards" when showing no influence in the test. The results are LOD scores on the hypothesis of a QTL being linked to the markers.

LOD scores arranged in the chromosome position and plotted are called Log-likelihood-ratio (LR) profiles and allows to asses visually the presence and number of potential QTLs. Yet, a LOD score threshold has to be established as sufficient to accept a locus as a potential QTL. Usually, a LOD score between 2 and 3 has been considered enough for single marker analysis (Churchill and Doerge, 1994). In the case of multiple marker analysis a permutation procedure is required to establish such threshold (Doerge and Churchill, 1996). Permutation tests consist in calculating multiple times, hundreds to thousands, the same analysis on a random resampling of the phenotypes. The procedure exchanges phenotypic values between individuals and calculates the distribution of LOD scores when no causal QTL-marker linkage exists. On this distribution, 90 and 95 percentiles are accepted such that only 10% and 5%, respectively, chance of false positive are allowed. This procedure can be done marker by marker or globally for the trait. The election in R/QTL implementation is to calculate a "traitwise" threshold to be applied to all markers, by pooling LOD values of all analysed markers.

In summary, MQM analyses the association between pair of markers with the phenotype accounting for the effect of other markers in the vicinity, returning a probability value for this association. The calculation of LOD threshold values per phenotype allows to accept or reject the hypothesis for association.

MQM analysis was applied to rosette shape descriptors to locate Quantitative trait loci related with rosette morphology. Phenotypes were measured at 5 consecutive days and a geometric (exponential) model was fitted for every shape descriptor time trajectory. The intercept and slope of the models were also used as input for MQM analysis.

3.2 Material and Methods

3.2.1 Biparental cross population

The biparental cross of Cape Verde Island and Argentat, France, accessions is one of the crosses performed, out of 6, at Ian Bancroft's laboratory (O'Neill et al., 2008). The crosses were initially designed for QTL mapping of seed lipids. The accessions for crossing were selected due to their diverse original location and variation for this trait. The Cvi-0 X Ag-0 cross was chosen for this experiment because the visual differences in rosette morphology between both parental accessions (see chapter 2).

RILs were generated by single seed descent for eight generations. The original population have 94 RILs genotyped at 71 Simple Sequence Repeat (SSR) and insertion/deletion (InDels) markers, mapped using Kosambi mapping function. The distribution of SNPs are 17,12 11,14 and 17 markers for chromosomes 1 to 5, as illustrated in the figure 3.1. In this experiment, 88 RILs were grown, due to individuals 3,7,38,44,49 and 54 did not germinate.

3.2.2 Experimental conditions

88 Recombinant Inbred Lines from the Cvi-0 x Ag-0 population were planted in a randomized schema at the NPPC facilities. 8 replicates of each RIL were placed in 36 trays containing 20 plants (5x4) each (figure 3.2 for an example). Seedlings were kept until flowering in PSI PlantScreen phenotyping device being watered automatically every day. Plant were also sent to the imaging chamber daily.

Seeds were sown on 24th June 2015 and kept in vernalisation for 7 weeks. On August 12th plants were pricked out to single pots and placed in trays in randomized blocks. On August 18th, trays were placed into PlantScreen device and maintained until 23rd August (5 days). Trays were imaged daily from the top. For compatibility with other experiments, all dates were translated to Days after Experiment started (DAE) instead of Days after Sowing.

The pots and greenhouse conditions were similar to the experiment described at chapter 2.

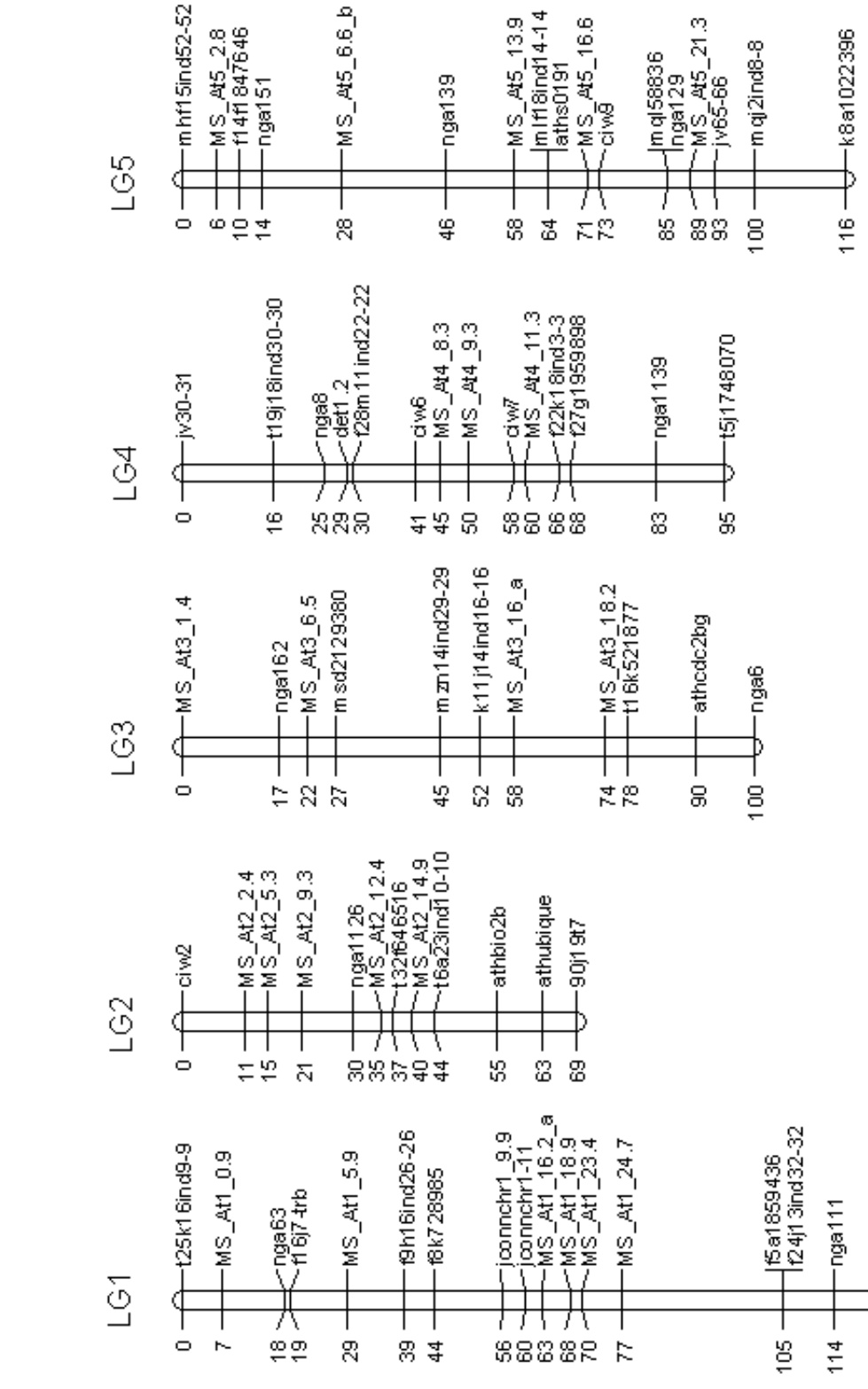


Figure 3.1: Genetic Map of markers in the experimental cross between Cvi-0 x Ag-0.
Reproduced from `https://www.jic.ac.uk/staff/ian-bancroft/arabidopsis-populations.htm\#CA`
71 SSR and InDels markers distributed along the 5 chromosomes in Arabidopsis thaliana outcross of Cvi-0 x Ag-0

Figure 3.2: Example of PSI PlantScreen tray with Cvi x Ag RILs population

3.2.3 Image processing and Shape Descriptors

PlantScreen internal software performs automatically the tasks of image correction, segmentation and Shape descriptor computing. Image processing pipeline is presented according to PSI personal communication. Figure 3.3 summarizes the workflow with an example.

PlantScreen camera objective is curved in a way that the whole tray is ensured to be in a single image. Therefore, the image is distorted with a moderate fish-eye effect and the software perform an initial correction on the image to "flatten" the picture and recover the original area of the objects.

The fish-eye corrected image is sent to the image segmentation pipeline. The first step is to overlap a predefined masking image with squares that delimits each pot. In consequence, the resulting image keeps only within-pots content, eliminating conveyor and tray parts. Every pot is now a so-called Region of Interest (ROI). A new gray-scale image is computed using the formula $I_{Gray} = 4 * I_{Green} - 3 * I_{Blue} - I_{Red}$ to each pixel (symbol I stands for Intensity) in each

ROI. On the new gray-scale image, a median filter is applied, i.e for each pixel, its neighbour pixels are gather and the median of these pixels becomes the new pixel value. The size of the neighbourhood is a set-up parameter, configured at 25 neighbours for these images.

Finally the gray-scale image can be thought as a collection of intensity values in a range from 0 to 255, whose frequency is arranged in a histogram. A threshold on pixel intensity is used to collect a binary image with values below and above it. Pixels above this threshold are classified as foreground and below it as background. The threshold was configured as 35% for these images and plant pixels are considered those classified as foreground. This binary image is itself also a mask that can be used to extract the colors in the original fish-eye corrected image, although this source of information is not used in this work.

Shape Descriptors

The Plantscreen device automatically calculates its own set of built-in shape descriptors described in Table 3.1 and illustrated in figure 3.4. The names and definitions of descriptors are based on PSI personal communication.

The shape descriptors calculated by PlantScreen software overlap partially with those from chapter 2. Rosette area and perimeter, Convex Hull area and perimeter, compactness, rosette roundness, convex hull roundness (called roundness2 in PlantScreen) and eccentricy are calculated by the same formulas. Rotational Mass Symmetry (RMS), Slender of Leaves (SOL) and Isotropy are specific of this software. Their formulas are described in table 3.1, and the figure 3.4 provides a visual guide to the elements required to calculate these shape descriptors, i.e Convex Hull, Skeleton, Circle with equal area than rosette.

RMS describes the uniformity of the rosettes. For rosettes that are compact, their area and convex hull area are close and a circle with same area as the plant will be near the border. Then the RMS will be small. On the contrary, loose rosettes has its pixels more spread, so the circle would have a radius smaller than the rosette radius, and many rosette pixels would be out of this circle. The RMS will be bigger.

SOL is a complicated measurement that seems to account for the extension of leaves from the centroid. The use of skeletons instead of perimeters reduce the error due to partially overlapping leaves.

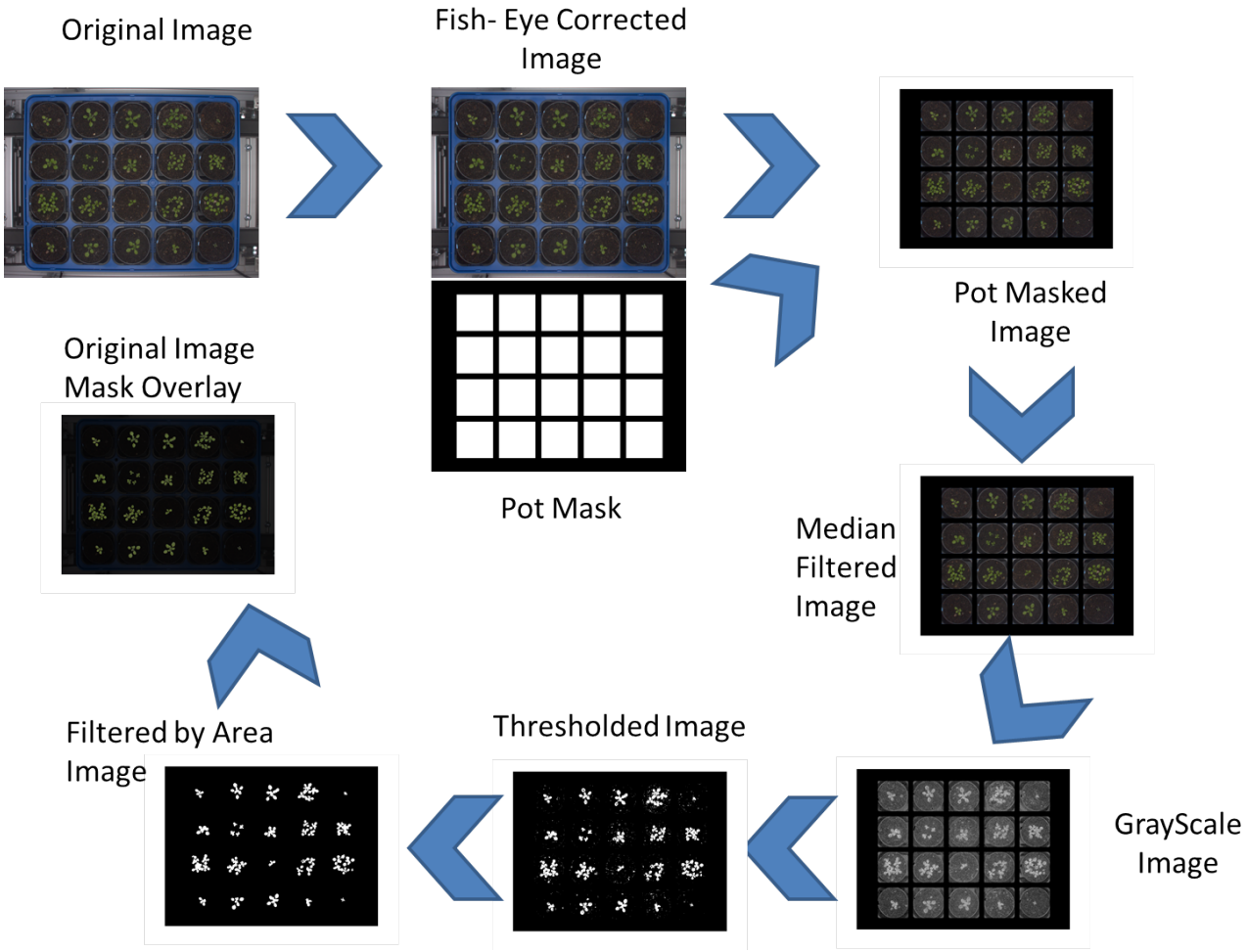


Figure 3.3: Representation of Image Processing Pipeline performed automatically by PSI PlantScreen

Descriptor	Units	Description	Computation	Notes
Area	Pixel or mm2	Area of plant Surface	Number of foreground pixels	Pixels to mm2 through a conversion rate
Perimeter	Pixel or mm	Length of plant perimeter	Number of border pixels	Pixels to mm through conversion rate
Compactness	Arbitrary Units	Ratio between rosette Area and Rosette Convex Hull	$\frac{Area_{Rosette}}{Area_{ConvexHull}}$	Convex Hull is the minimum convex (no lines between two points exits the shape) polygon surrounding the rosette pixels 0 - Empty object – 1 – Filled object
Roundness	Arbitrary Units	Ratio between Rosette Area and squared Rosette Perimeter	$\frac{4 \cdot \pi \cdot Area}{Perimeter^2}$	0 –Line 1 – Circle
Roundness 2	Arbitrary Units	Ratio between Convex Hull Area and Convex Hull Perimeter	$\frac{4 \cdot \pi \cdot CH_{Area}}{CH^2_{Perimeter}}$	0 –Line 1 – Circle

Table 3.1: Shape Descriptors as Defined in PlantScreen Software

. . . continued

Descriptor	Units	Description	Computation	Notes
Eccentricity	Arbitrary Units	Ratio of distance between foci of ellipse and its major axis length		The ellipse is calculated as having the same second-moments than rosette point distribution
Rotational Mass Symmetry(RMS)	Arbitrary Units	Ratio between Convex Hull Areas outside and inside of a circle with same area as plant.		Circle centre is at plant centroid. The radius is weighted with Compactness
Slender of Leaves (SOL)	Arbitrary Units	Square length of plant skeleton divide by Area	$\frac{Perimeter^2_{Skeleton}}{Area_{Rosette}}$	
Isotropy	Arbitrary Units	Roundness of the polygon formed from leaf tops	$\frac{4 \cdot \pi \cdot Polygon_{Area}}{Polygon^2_{Perimeter}}$	Similar to Roundness but more robust.

Table 3.1: Shape Descriptors (Continued)

Figure 3.4: Example of elements to calculate Shape Descriptors. Area: Dark Green pixels inside rosette at left and right images. Perimeter: Red pixels surrounding rosette at left image. Convex Hull. All coloured pixel (non-black) at left image. Regions of Convex Hull No intersecting with rosette are in blue, light blue, light green, orange and yellow. Skeleton: Blue pixels inside the rosette at left and right images. Circle with same area than Rosette. Red circle in right image. Polygon joining leaf tips not represented in these images

Isotropy is a convex hull roundness-like metric, but instead of a convex polygon, it uses a concave one that touch every leaf tip. It is claimed to be more robust, since segmentation artefacts are not playing an important role in its error.

3.2.4 QTL mapping

QTL mapping was initially performed considering every Shape descriptor at each single day, so that for example, AREA_MM become five traits AREA_MM_0, AREA_MM_1, AREA_MM_2, etc. where numbers represent Days After Experiment started (DAE).

In addition, to take advantage of Shape Descriptor time course monotonicity, and given the number of replicates per RIL (8 plants), a geometrical model has been applied to each RIL for each parameter. The geometrical growth model (equation 3.1) allows to model exponential growth, exponential decay and constant slope, i.e. null, and it was a suitable approach for all traits. The original growth model (equation 3.2) uses N_0 as population size at time zero, and for consistency with non-growth models we rename it as "A". The growth rate is usually named r and correspond to the slope, so we renamed it as "B" also for consistency (see equation 3.3).

$$dN/dt = rN \tag{3.1}$$

$$N_t = N_0 \cdot e^{r \cdot t} \tag{3.2}$$

$$Shape_t = A \cdot e^{B \cdot t} \tag{3.3}$$

All data handling, plotting and analysis has been performed in the statistical software package R. Geometric models were fitted using non-linear squares with the function *nls*. QTL mapping has been performed with the package R/QTL. Multiple QTL Mapping has been done with the function *mqmscanall*. The automatic cofactor selection used the *mqmautocofactors* function, choosing a maximum of 50 cofactors to test. Permutation tests have been performed for each trait_DAE (Days after Experiment started) using the function *mqmpermutation* with 500 permutations and including cofactors. All the analysis were performed after coercing the cross to be "riself", meaning RILs by selfing, and augmenting the data with the function *mqmaugment* with the option "minprob" set up as 1.0. MQM scans, auto-cofactors and permutation used the augmented, without missing data, version of the cross. After results were calculated, pseudo-markers were eliminated from the markers set, keeping only the original genotyped markers.

3.3 Results

3.3.1 Phenotypic variation

All plants were onto the PlantScreen Device for 5 days (typed as "Day After Experiment Start", DAE, from 0 to 4). Figure 3.5 show 25 plants at DAE 2. They are the 8 replicates of parental accessions plus RILs CA83 and CA16. The parental Cvi had a low germination rate, so only one replicate was included. This does not affect any analysis in this chapter since parentals are included for visual comparison purposes. Figure 3.6 represent the rankings for every plant according to Compactness at DAE 2. Table 3.2 present the values of Compactness and the rank of every plant in the figure 3.5. For simplicity, the rank in the population (Global Rank) has been mapped to a rank for the plant in the table (Table Rank). The plants in the figure

3.2 are organized as a parental/RIL per row, and within row plants are sorted by rank.

The two figures show that RILs and Ag-0 had different rosette structures between strains and consistent within RIL/Parental. Cvi-0 replicate is clearly the most compact, and Ag-0 and CA16 are more similar between them than to CA85, which is clearly reflected in the ranking.

For a more comprehensive view of Shape Descriptors, figure 3.7 shows histograms for each one at DAE 0. Most parameters were normally distributed with a variable degree of skewness. Although the parentals were chosen due to their different rosette shape, the F8 RILs still present transgressive segregation. That is, the distribution of RILs' rosette shape quantitative traits expand in the population to values that range beyond those of their parentals, but keeping close to normal distribution rather than a uniform one. For that reason, the parentals were either close to the mean value or were together near one of the extremes. For example, Eccentricity and Roundness2 had both parentals in one of the extremes. Area, Perimeter, Roundness, Rotational Mass Symmetry and Slender of Leaves were close to the average. For Compactness, Ag-0 is in the middle while Cvi-0 is in one extreme and the same occur in Isotropy.

Figure 3.8 present the average per RIL developmental trajectories throughout days,split by shape descriptors. Rosette area and perimeter show that plants were growing exponentially while the other morphological parameters had a monotonic, either positive or negative, trajectory alongside growth. Eccentricity and Roundness-2 indicate plants were getting more round shaped across time arriving to a steady-state around the 2nd or 3rd day. However, Isotropy and Roundness does not present any direction over time, being Isotropy a very "disperse" descriptor. Slender of Leaves and Compactness show that most plant were getting a more dense habit along days.

In general, most accessions were growing exponentially, extending leaves and petioles, at the same time covering gaps between leaves, therefore increasing Compactness and Slender Of Leaves, and becoming more round-shaped. To demonstrate visually the rosette development figures 3.9 and 3.10 and table 3.3, display four individual plant examples, the two parentals, CA16 and CA83, across time. The figures and the table illustrate the relationship between growth and shape throughout time.

Cvi-0 and CA83 were bigger than the other accessions, in terms of area, than Ag-0 and CA16. However, The perimeter of Cvi-0 is in the range of Ag-0 and CA16 while CA83 has

Figure 3.5: Example of 4 varieties from the Biparental cross, parentals, Ag-0 and Cvi-0, and RILs, CA83 and CA16, sorted by ranking according to Compactness at DAE 2 (figure 3.6 and table 3.2).
In each row only one accession is represented. Only one replicate from Cvi was available.

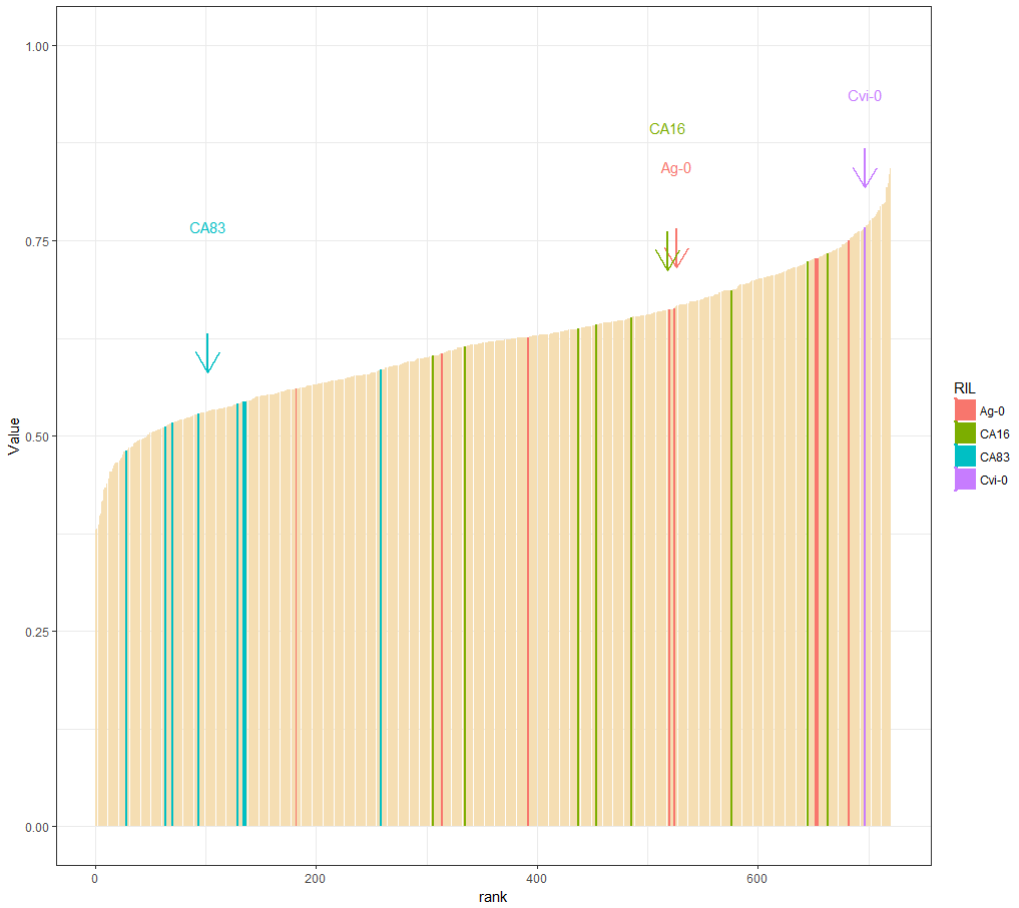


Figure 3.6: Ranking of the Biparental cross population by Compactness at DAE 2. Ranking calculated over the individuals. Colours are orange = Ag-0;Green = Ag-16;Blue = CA83;Purple = Cvi-0. Each bar correspond to an individual. Average value is indicated with an arrow

RIL	Tray	Position	DAE	Compactness	Global Rank	Table Rank
Ag-0	33	C3	2 days	0,56	182	8
Ag-0	38	A1	2 days	0,61	314	11
Ag-0	51	A5	2 days	0,63	392	13
Ag-0	46	D2	2 days	0,66	520	17
Ag-0	60	A1	2 days	0,66	524	18
Ag-0	42	B4	2 days	0,73	652	21
Ag-0	29	A5	2 days	0,73	654	22
Ag-0	55	C3	2 days	0,75	682	24
CA16	29	A1	2 days	0,60	306	10
CA16	37	D2	2 days	0,61	335	12
CA16	42	A5	2 days	0,64	437	14
CA16	46	C3	2 days	0,64	454	15
CA16	51	A1	2 days	0,65	485	16
CA16	55	B4	2 days	0,69	576	19
CA16	59	D2	2 days	0,72	645	20
CA16	33	B4	2 days	0,73	663	23
CA83	33	D3	2 days	0,48	28	1
CA83	38	B1	2 days	0,51	64	2
CA83	47	A2	2 days	0,52	70	3
CA83	29	B5	2 days	0,53	94	4
CA83	60	B1	2 days	0,54	129	5
CA83	42	C4	2 days	0,54	134	6
CA83	51	B5	2 days	0,54	136	7
CA83	55	D3	2 days	0,58	259	9
Cvi-0	26	A1	2 days	0,77	697	25

Table 3.2: Example of rosettes from Biparental cross population at DAE 2. Population parentals, Ag-0 and Cvi-0, and RILs, CA83 and CA16. Table contain Compactess values at DAE 2 and ranking position considering the whole population (Global rank), and within rosettes in this table (Table rank) for comparison. Individual identifiers are the combination of Trays and Position in the tray (letter for rows and numbers for columns). There was 8 replicates of each RIL and Ag-0, but only one of Cvi-0. The table is sorted first by Accession and later by Global Rank

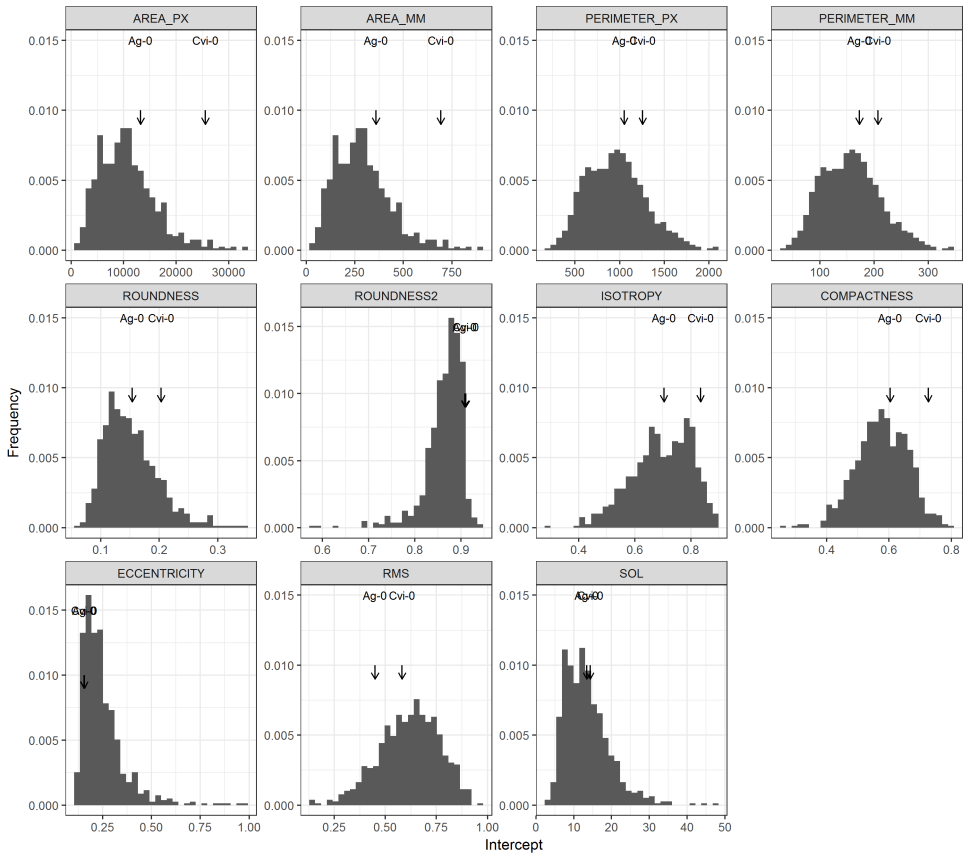


Figure 3.7: Histogram showing average value per RIL and Shape Descriptor at DAE = 0. Two black arrows show the average value for parental Cvi-0 and Ag-0

bigger values. The reason is the longer petiole length in CA83 than in the other three, having this plant a looser habit as observed in figure 3.10 in the values of Compactness and Slender of Leaves.

As it has been stated in chapter 2, shape descriptors measure multiple aspects of rosette architecture at once and it has been explained that mutual information is shared between them, e.g compacness and Slender of Leaves. Thus, it is expected a degree of correlation between descriptors value that facilitates to split descriptors in groups. Figure 3.11 shows the correlation between pairs of descriptors. Eccentricity and Roundness-2 (Convex Hull roundness in chapter 2) had a correlation of -0.88 indicating that both are good measurements for circularity, however, a deeper study on shape descriptors mathematics (not included in this thesis) indicates that eccentricity is more robust to segmentation artefacts than Roundness-2. Hipothetically, we expect high correlation between roundness, roundness2, eccentricity and isotropy, due to their similar definition, but roundness do not correlate with any of them. Again, the experience suggest that roundness, being dependent of perimeter, is very variable parameter very affected

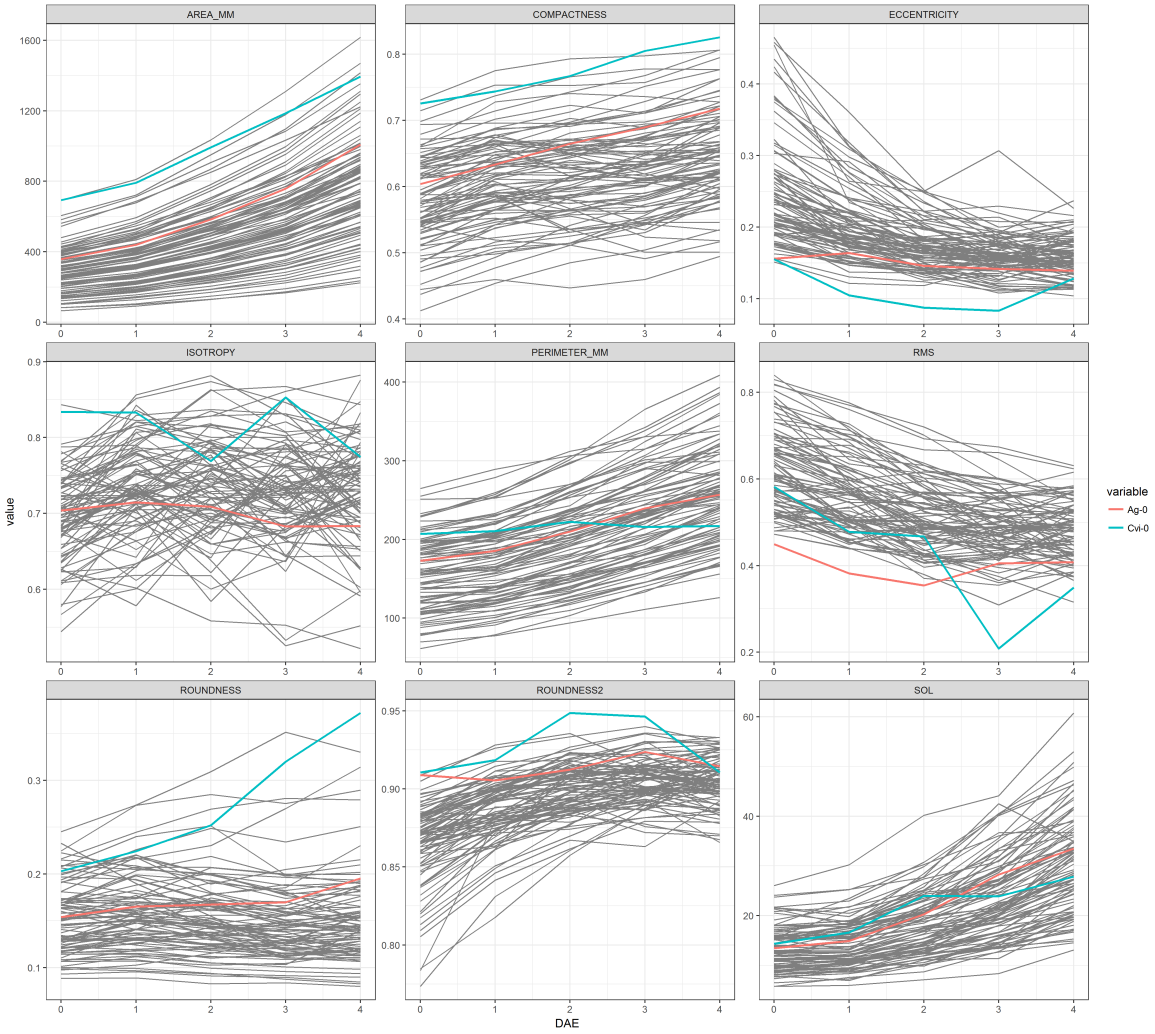


Figure 3.8: Shape Descriptors Time Trajectory. Each line represent the RILs average per day along the 5 days the experiment lasted. For comparison, the parental Cvi-0 is coloured in red and Ag-0 in blue

Figure 3.9: Example of (top down) Cvi-0, Ag-0, CA83 and CA16 along DAE 0 to 4

for leaf movement and overlapping. For this reason, it correlates with compactness at the end of vegetative growth, when the plants cover most of its space and perimeter counts more for the outer perimeter of the rosette than the leaf sides and petioles. Area and perimeter correlates due to both are measuring rosette size. For bigger leaves, the more area and more perimeter they have. Their correlation is 0.76 indicating as well that for a certain area exist variation in perimeter according to leaf shape, e.g rounded leaves correspond to smaller perimeters than elongated ones. Slender of leave correlates with perimeter due to the skeleton length is a robust measure of leaf length in the same way than perimeter is a measure of leaf outer surface length.

The general idea is that shape descriptors can be interpreted as one single multivariate descriptor, so that all of them represent the rosette global architecture. For that reason QTLs associated to any of these descriptors may be considered as QTL for rosette architecture rather than separated them in QTL for area, QTL for roundness, etc.

Each shape descriptor trajectory was modelled using equation 3.3. Parameters A and B, i.e intercept and slope, were fitted using non-linear least squares. After modelling the curves, the time trajectory get condensed into these two parameters. Intercept would be the value of the parameter at time 0, so the initial averaged value, and the slope would be the direction of the shape descriptor when growing, indicating the speed of change.

Histograms, figure 3.12, of Intercept values show that they are normally distributed, while the position of parentals at each value are similar to the Shape at day 0. For the slope, every shape descriptor becomes normally distributed and potentially interesting variation emerge. For example Isotropy show values between [-0.5, 0.5] so some RILs decay in isotropy while others are growing. Other example is compactness, with an Ril with an slope of ~ 0.075 indicate that quickly cover its region, while other RIL reduce its compactness at $\sim$ -0.05, so that it almost does not vary its coverage.

Broad sense heritability was calculated from a mixed model using RILs names as random effect, so that Genetic variance is calculated as the random effects variance, and environmental variance from the model error. The heritability for most traits and DAE combinations were over 20%, with the exception of Eccentricity and Isotropy at DAE 3 and RMS at DAE 4. Area, Perimeter, Compactness and Roundness had the largest heritability, around 67% for Area and all DAEs between 57 and 67 for Roundness across time. In general, heritability has not heavy

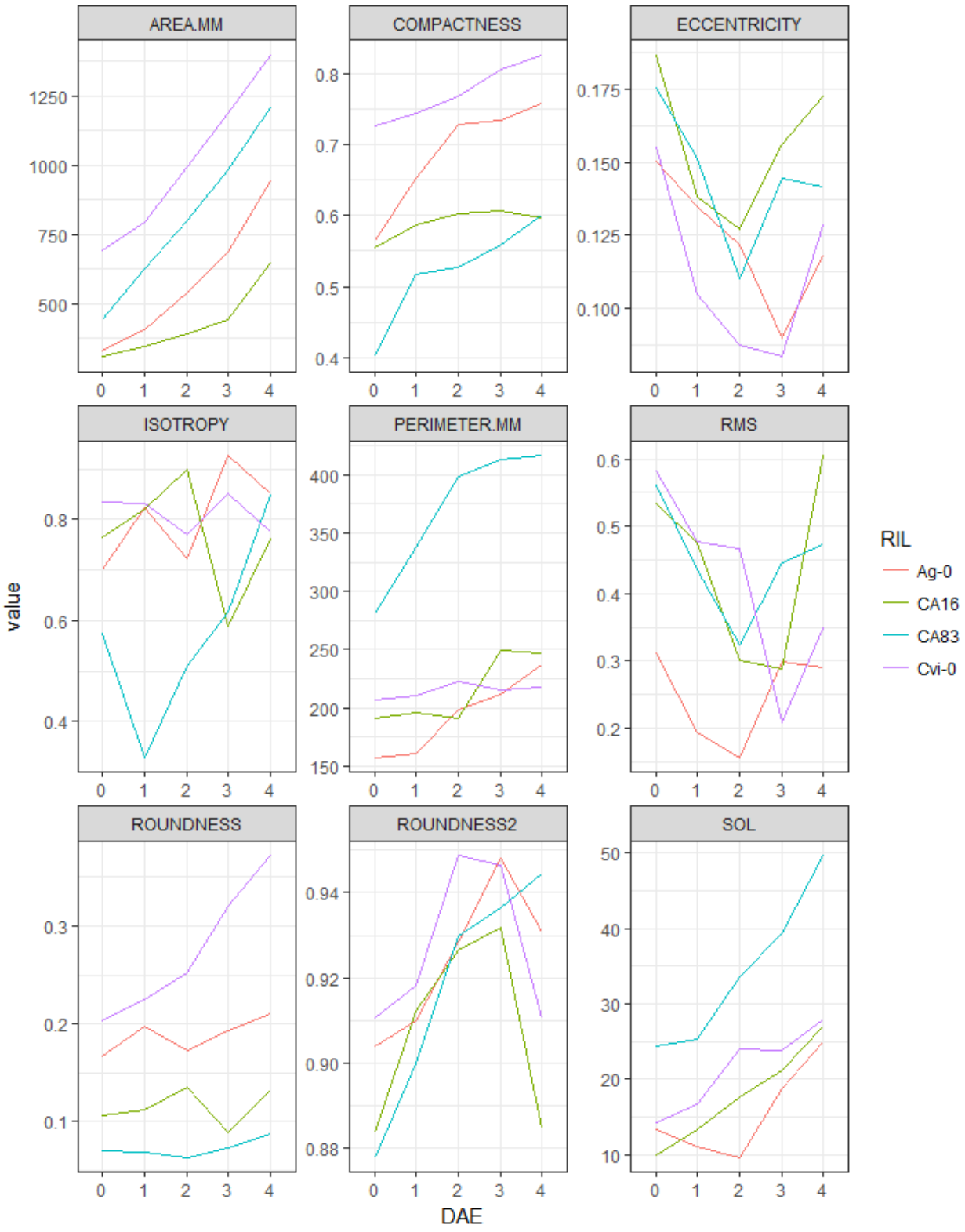


Figure 3.10: Time course trajectory for the selected example at figure 3.9

Table 3.3: Phenotypic values for DAE 0 to 4 for the selected example at figure 3.9

AREA.MM			DAE		
	0	1	2	3	4
Cvi-0	692,44	791,66	992,43	1185,94	1395,28
Ag-0	331,42	404,82	538,29	688,1	941,95
CA83	440,83	624,78	797,84	984,09	1207,16
CA16	307,28	345,41	393,28	441,86	643,48

COMPACTNESS			DAE		
	0	1	2	3	4
Cvi-0	0,73	0,74	0,77	0,81	0,83
Ag-0	0,56	0,65	0,73	0,73	0,76
CA83	0,4	0,52	0,53	0,56	0,6
CA16	0,55	0,59	0,6	0,61	0,6

ECCENTRICITY			DAE		
	0	1	2	3	4
Cvi-0	0,16	0,1	0,09	0,08	0,13
Ag-0	0,15	0,13	0,12	0,09	0,12
CA83	0,18	0,15	0,11	0,14	0,14
CA16	0,19	0,14	0,13	0,16	0,17

ISOTROPY			DAE		
	0	1	2	3	4
Cvi-0	0,83	0,83	0,77	0,85	0,77
Ag-0	0,7	0,82	0,72	0,93	0,85
CA83	0,58	0,33	0,51	0,62	0,85
CA16	0,76	0,82	0,9	0,59	0,76

PERIMETER.MM			DAE		
	0	1	2	3	4
Cvi-0	207,1	210,54	222,52	215,82	217,15
Ag-0	157,69	160,62	197,95	211,71	237,01
CA83	280,3	337,38	398,22	412,42	416,31
CA16	191,38	196,5	191,03	249,85	247,4

RMS			DAE		
	0	1	2	3	4
Cvi-0	0,58	0,48	0,47	0,21	0,35
Ag-0	0,31	0,19	0,16	0,3	0,29
CA83	0,56	0,44	0,32	0,45	0,47
CA16	0,53	0,47	0,3	0,29	0,61

ROUNDNESS			DAE		
	0	1	2	3	4
Cvi-0	0,2	0,22	0,25	0,32	0,37
Ag-0	0,17	0,2	0,17	0,19	0,21
CA83	0,07	0,07	0,06	0,07	0,09
CA16	0,11	0,11	0,14	0,09	0,13

ROUNDNESS2			DAE		
	0	1	2	3	4
Cvi-0	0,91	0,92	0,95	0,95	0,91
Ag-0	0,9	0,91	0,93	0,95	0,93
CA83	0,88	0,9	0,93	0,94	0,94
CA16	0,88	0,91	0,93	0,93	0,88

SOL			DAE		
	0	1	2	3	4
Cvi-0	14,32	16,63	23,97	23,86	27,87
Ag-0	13,28	11,14	9,66	18,69	24,93
CA83	24,32	25,25	33,49	39,32	49,63
CA16	10,01	13,38	17,57	21,2	27,02

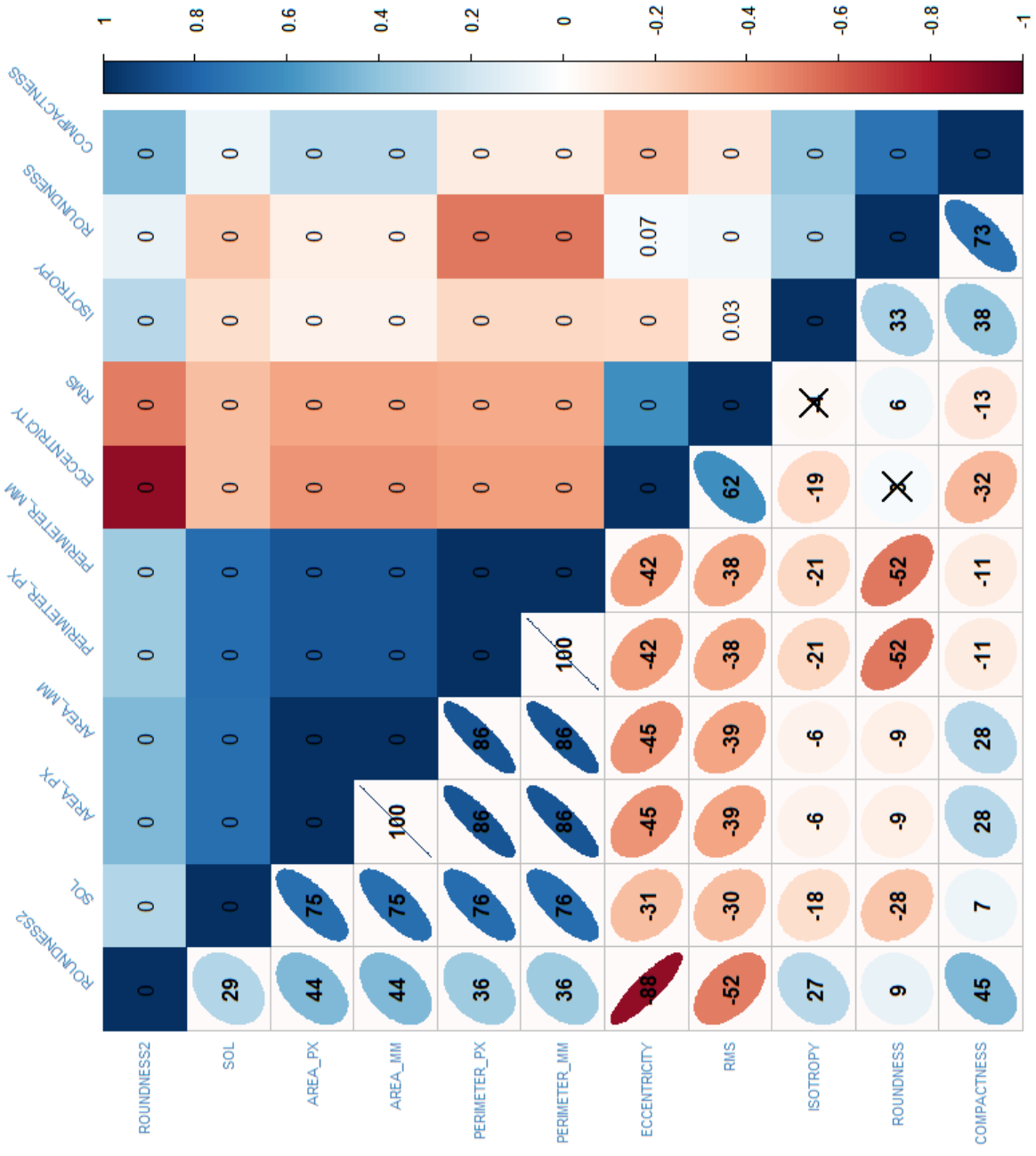


Figure 3.11: Pairwise correlation between descriptors values. Blue represents high positive correlation and red high negative one.

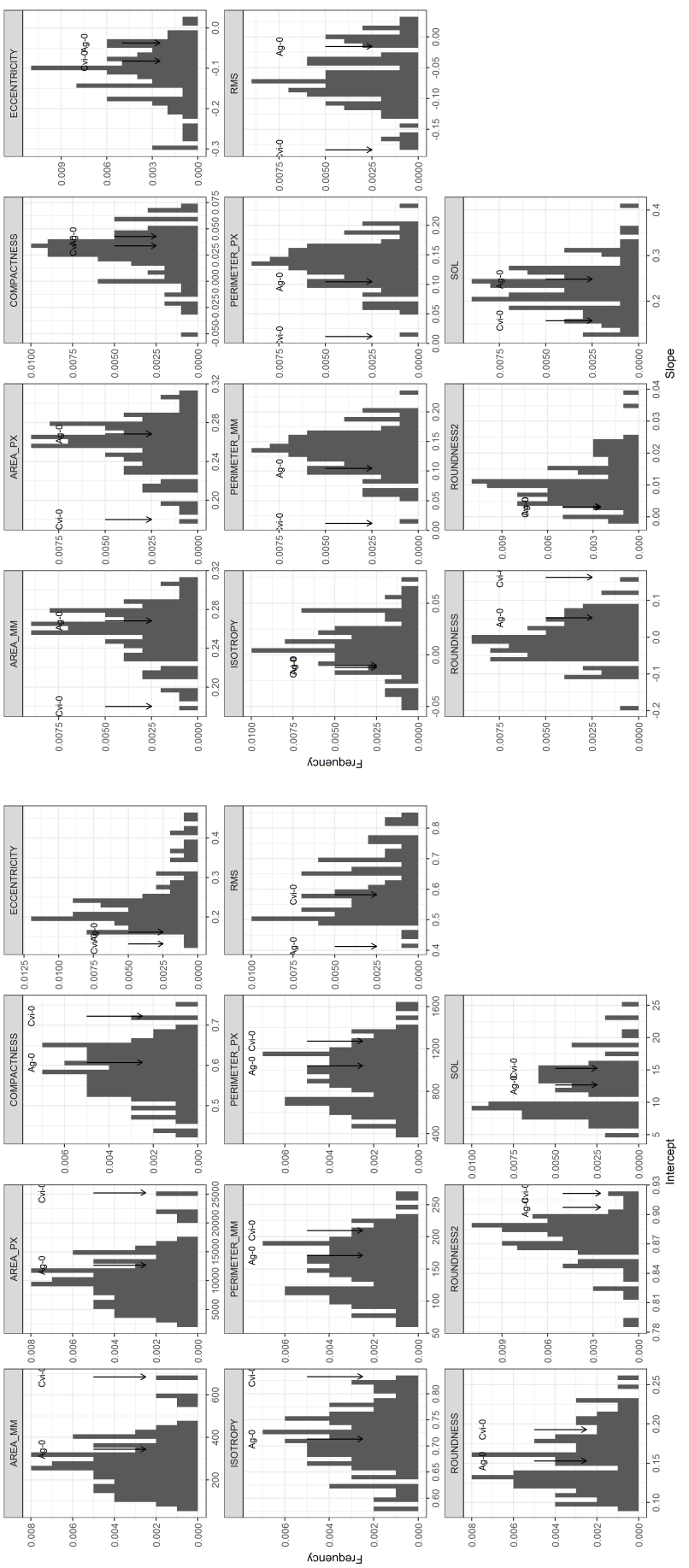


Figure 3.12: Histogram for Intercept and Slope of geometric model fit to each descriptor and RIL. Black arrows show Intercept and Slope of parentals Cvi-0 and Ag-0

fluctuations for any descriptor in time, but RMS, Solidity and Eccentricity have seen reduced their heritabilities from DAE 0 to 4.

3.3.2 Multiple QTL mapping

Multiple QTL Mapping is a variation of the Composite Interval Mapping, that allows the use of several markers as cofactors. The advantage is that, while testing markers for significance in its association with a phenotype, variation in the trait due to variation in markers other than tested one is controlled. The package R/qtl has the function *mqmautocofactors* to automatically, based on markers density and backward elimination procedure, select a sparse set of cofactor markers. The immediate effect is a much cleaner QTL profile, with less noise and more adequate LOD values per markers. The application of MQM to each single day and shape descriptor average RILs values return QTL profiles plot. The profile plots represent the Log of Odds (LOD) score for each marker, sorted by chromosome and position within. By overlapping QTL profiles of different DAEs, an insight of how much influence a possible QTL is getting, or losing, through time. Finally, permutation test helps to find the LOD score threshold to accept a markers as pinning a QTL region.

The mapping population has 89 RILs genotyped for 71 markers (17, 12, 11, 14 and 17 for chromosomes 1 to 5). 97.3% of markers had been genotyped, having a 39% AA, 58.5% BB and 2.1% AB (heterozygous) (see table 3.4). There were 181 missing genotypes, in average 2.55 RILs were not genotyped for each markers. The cross were transformed, from the default "F2" in the software, to "RILs by Selfing", which eliminates heterozygous markers. This is required to ensure correct genotypic probabilities are assigned during the Gaussian Mixture modelling. For MQM analysis, no missing data is accepted in the routine, so a procedure called genotype augmentation was performed. Using the minimum probability parameter as 1.0 all missing genotypes were imputed as the most probable value according to their neighbouring markers. The result is a population homozygous for all markers and without missing data (figure 3.14).

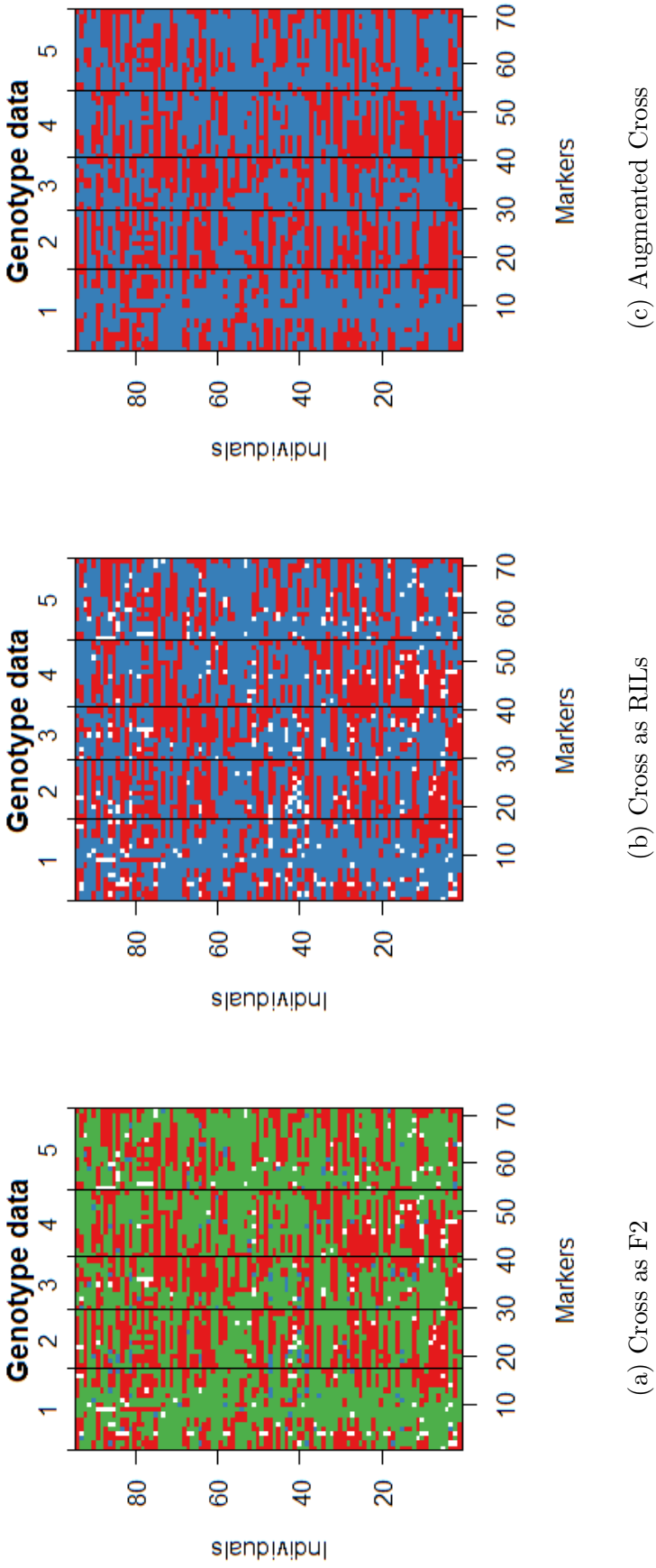
Map distances need to be recalculated after population modifications. Figure 3.15 shows that the recalculation produce a map expansion respect to the one at figure 3.1, being the genetic distance between markers increased. The average distance between markers is 7.64cM, in a range of 504cM for the whole genome, the maximum distance between two adjacent markers

Descriptor\DAE	0	1	2	3	4
AREA_MM	68.68%	68.87%	67.99%	67.47%	66.95%
COMPACTNESS	56.99%	59.24%	64.02%	65.54%	67.70%
ECCENTRICITY	36.45%	28.43%	20.01%	15.46%	24.29%
ISOTROPY	16.58%	21.45%	21.36%	15.43%	16.50%
PERIMETER_MM	68.24%	67.78%	65.25%	62.63%	59.89%
RMS	36.86%	31.82%	26.54%	23.21%	17.44%
ROUNDNESS	57.14%	59.18%	60.02%	61.63%	61.84%
ROUNDNESS2	32.98%	27.04%	26.39%	25.43%	31.86%
SOL	50.38%	43.43%	36.92%	30.86%	25.03%

Figure 3.13: Heritability of Shape Descriptors per DAE. Biparental Population Cvi-0 x Ag-0

Chromosome	n.mar	length	ave.spacing	max.spacing
1	17.00	124.01	7.75	28.37
2	12.00	69.06	6.28	10.93
3	11.00	99.94	9.99	17.75
4	14.00	95.28	7.33	16.44
5	17.00	116.20	7.26	17.28
overall	71.00	504.48	7.64	28.37

Table 3.4: Cvi-0 x Ag-0 population genetic map characteristics



(a) Cross as F2 (b) Cross as RILs (c) Augmented Cross

Figure 3.14: Genotypic values for Cvi-0 x Ag-0 population.
a). Cross as an "F2". Genotypes AA = Red,BB = Green, AB =Blue,missing = White.
b). Cross as "RILs by Selfing". Genotypes AA = Red, BB = Blue, missing = White.
c). Missing data imputed according to neighour markers recombination frequencies

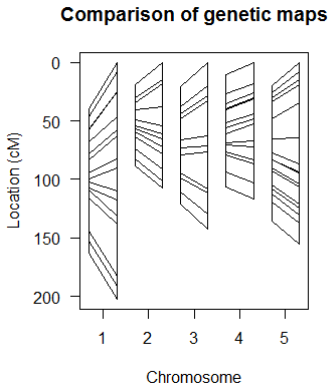


Figure 3.15: Genetic Map corresponding to SNP markers genotyped for Cvi-0 x Ag-0 RILs. The absence of 4 RILs generate a map expansion. At each chromosome the original map is at left side, and the calculated map at right.

was 28.37 cM.

As stated before, MQM results are essentially LOD scores supporting, or not, the hypothesis of having a QTL close to genotyped markers associated to phenotypic variation. To comprehend genome-wide LOD scores they are usually represented as Log-likelihood (LR) profile plots. These consist in markers sorted by chromosome position along the X axis and their LOD scores on the Y axis. Hill-like peaks on the LR profile indicates the location of a potential QTL. In this experiment 66 LR profiles, for 55 phenotypes corresponding to Shape_DAE and 11 phenotypes corresponding to Shape_Intercept and Shape_Slope, are produced. The whole collection of plots are difficult to interpret when put together. Thus, the approach here is to show QTL profiles for each descriptor separately, but keeping the different DAE in the same plot (figure 3.16). Two different plots were produced for each intercept and slope (figures 3.17 and 3.18)).

From these plots potential QTLs were extracted in Chromosome 1 for Area, SOL and Eccentricity. In Chromosome 3, at the rightmost section for Roundness and compactness, and at the left for RMS, Isotropy and SOL. Both seems to be found also for Roundness2. In Chromosome 2 there are low LOD potential candidates at the left and the right. Finally Chromosome 4 seems to harbour a potential QTL in the middle of the chromosome for Roundness, SOL and Perimeter.

These plots also address the relevance of measuring morphological traits across time, since the significance of association for each possible QTL may change in different days.

The kind of plots that help together to compare potential QTLs found by different descriptors are the raster plots stack all Descriptors and DAE into one single plot and map the LOD

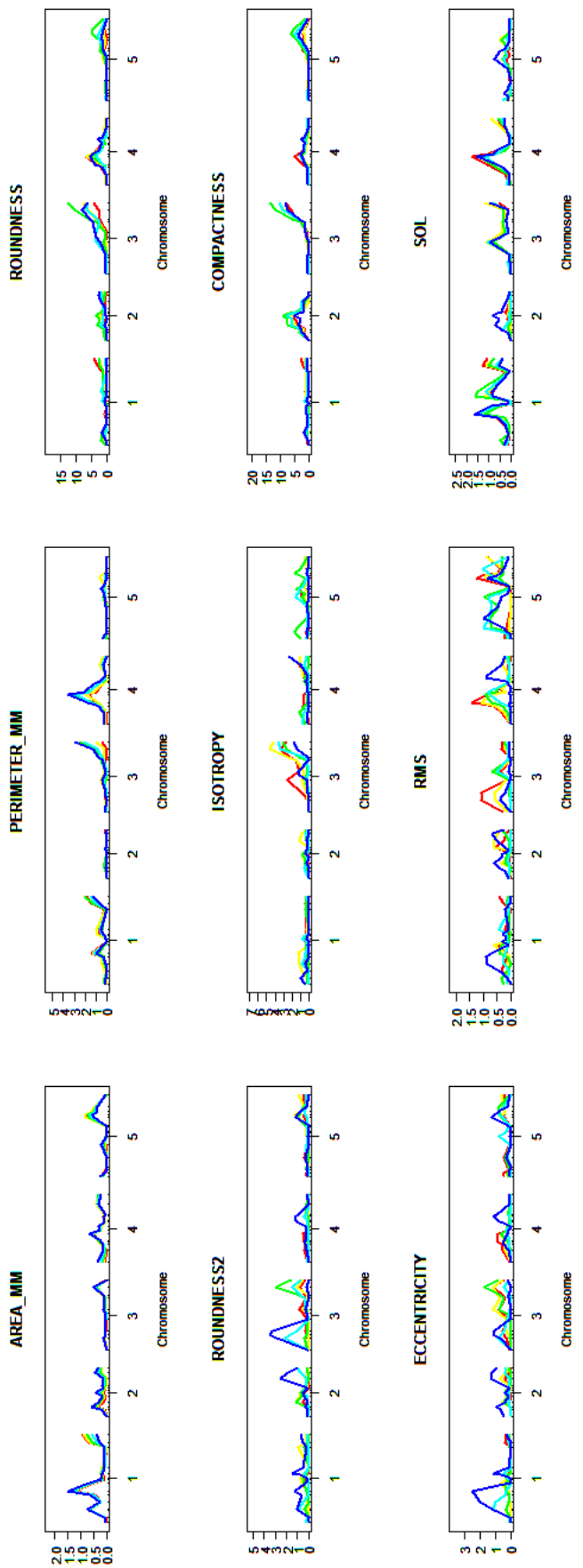


Figure 3.16: QTL profile plots for Descriptors_DAE phenotypes. At each plot a single descriptor at five DAE are represented.

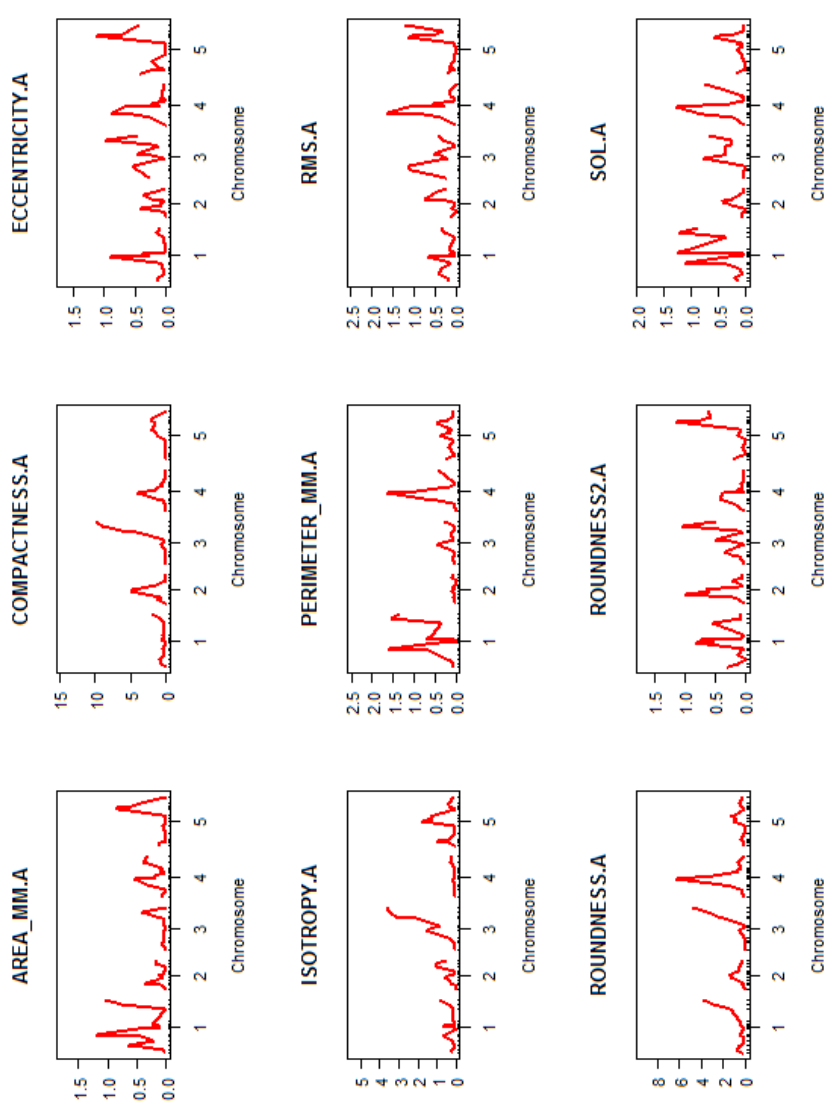


Figure 3.17: QTL profile plots for Descriptors Intercept phenotypes. At each plot a single descriptors' Intercept are represented.

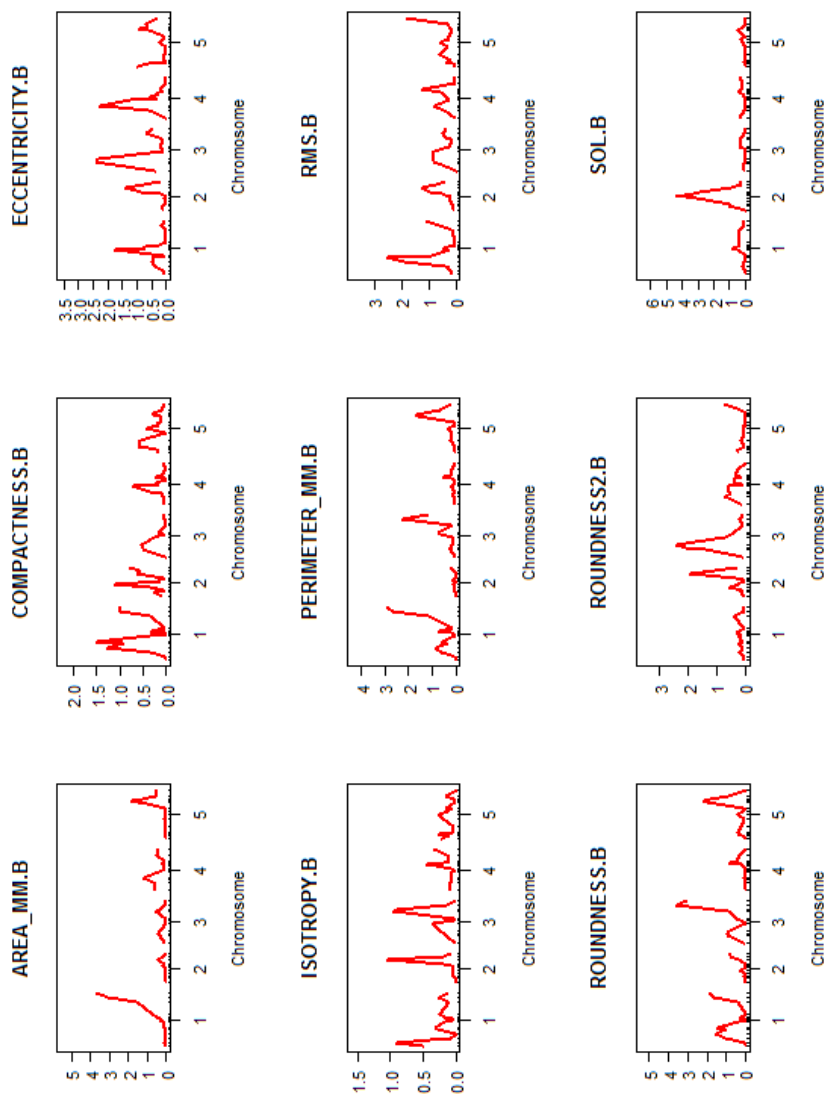


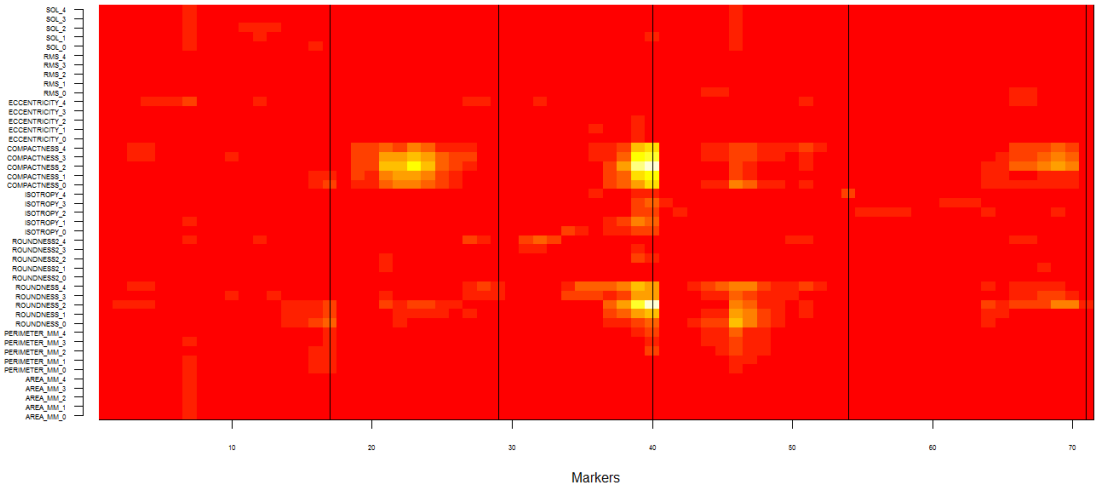
Figure 3.18: QTL profile plots for Descriptors Slope phenotypes. At each plot a single descriptors' Slope are represented.

height to a colormap. A joint LR profile has been added under it to facilitate the comparison of the colormap and the height. Unfortunately R/QTL raster plot do notaccount for the genetic distance in the X axis, but rather place each marker equidistantly. A third plot that aids visualizing the genetic distance between potential QTLs are the circular maps, where the chromosomes are draw as a circumference, with every markers significant at the MQM analysis is plotted at their genetic distance and lines connecting markers significant for the same trait pass through a symbol in the center, having a different color per trait. These three kinds of plots are presented as single figures for groups of descriptors (figure 3.19 for Descriptors and DAE, and 3.20 for Descriptors Intercept and Slope)

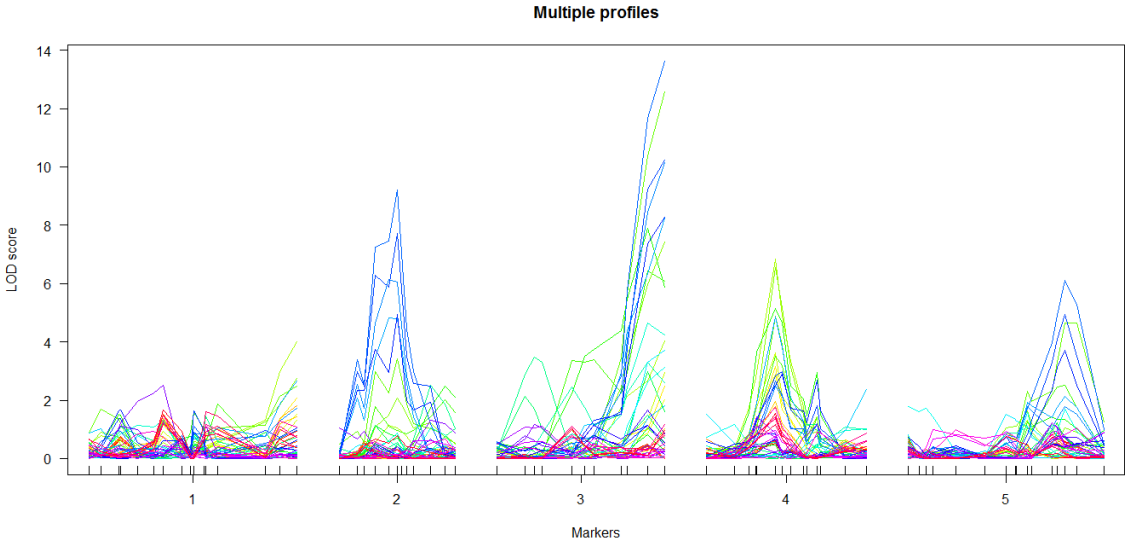
LOD thresholds calculated from permutation procedures described at table are used to identify markers significantly associated to shape descriptors variation. Table 3.5 (for the combination Shape Descriptors - DAE) contains only those markers with an "effect size" over the statistical significance level of 10 or 5%. It suggest a strong support for two "Shape QTLs". The first one is at the center of chromosome 2, surrounding the marker MS_At2_9 at 21.3 cM and MS_At2_1 at 34.88cM. It was found for compactness at DAE 0 to 4, with LOD values over 5, maximum at 9.20, being the LOD threshold close to 3. Roundness provide some support for this QTL, although the LOD scores are lower, between 2 and 4, but the calculated threshold is between 2.48 and 3.22. The second Shape QTL with strong support is at the end of the chromosome 3, between 89.5 and 99.94 cM . High LOD scores for this QTL is found at Roundness at DAE 3 (LOD =12.57), and for Compactness DAE 2 and 3 (max LOD = 12.63). Isotropy, Roundness and Roundness-2 also show peaks in this QTL but closer to the 5% threshold. Some other peaks suggest possible QTLs at Chromosome 1 for Area, Eccentricity and Roundness2 but these are not significant. In the middle of Chromosome 4, Solidity and Perimeter indicates another possible QTL, but still under threshold.

MQM on the intercept and slopes of geometrical models support the QTLs found at day by day analysis, but in general the signal is weaker, the maximum LOD peak is 10 and after this 6, a more sparse group of peaks appears in the image 3.20, although they were not crossing the 10% threshold. These results are available at table 3.6.

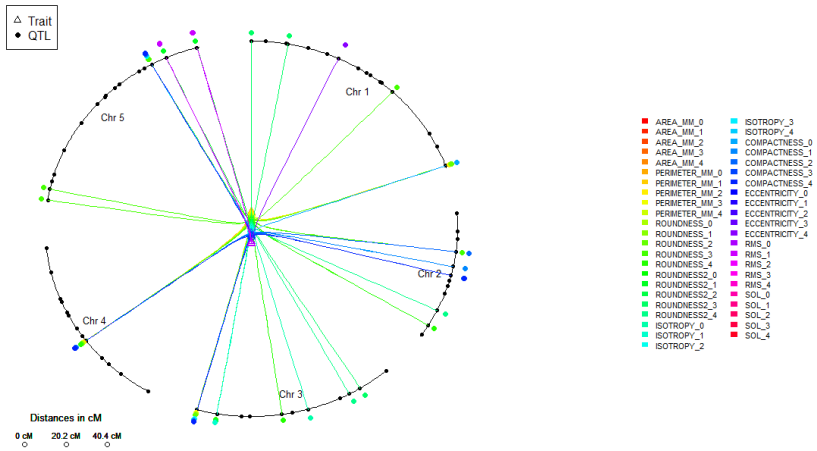
From the table it is observed that the QTL at middle of chromosome 2 has his highest LOD at Compactness Intercept (4.58 with 5% threshold of 2.90) and Solidity Slope (LOD = 4.43



(a) Raster image showing MQM results for all combinations of Shape Descriptors and DAE.

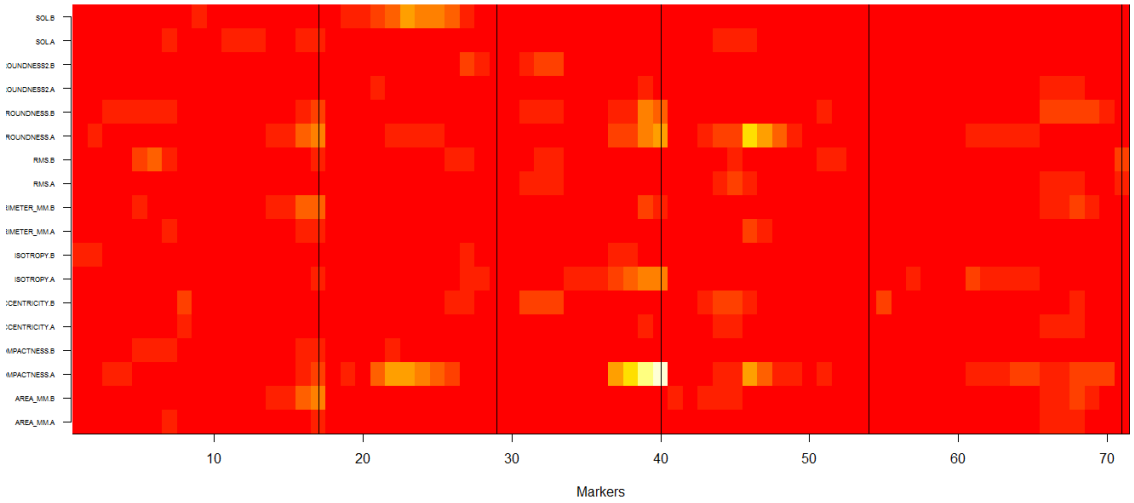


(b) QTL profile for all Descriptors and DAE pooled

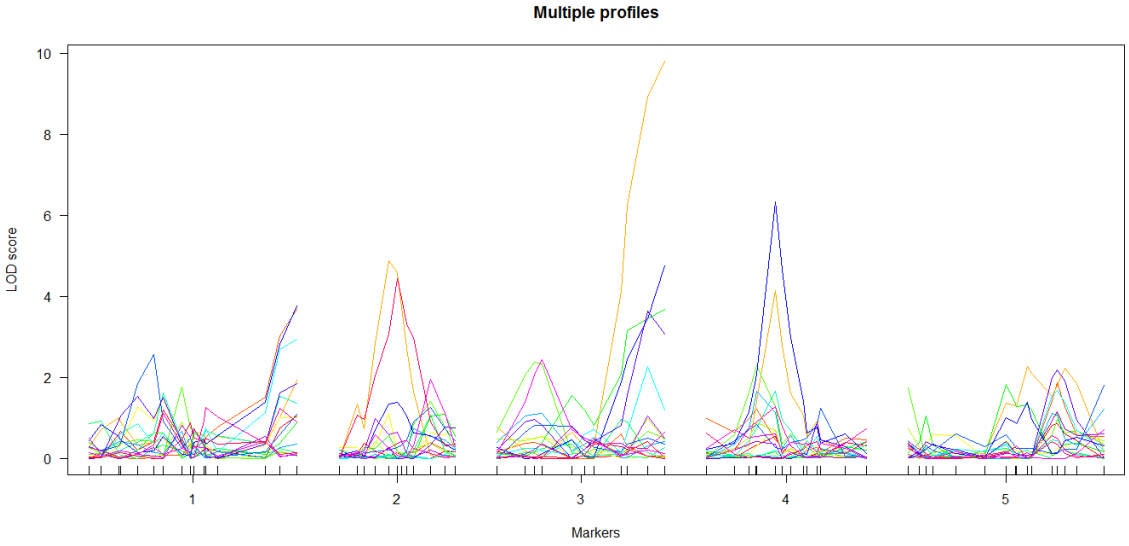


(c) Circle plot for all Descriptors and DAE pooled

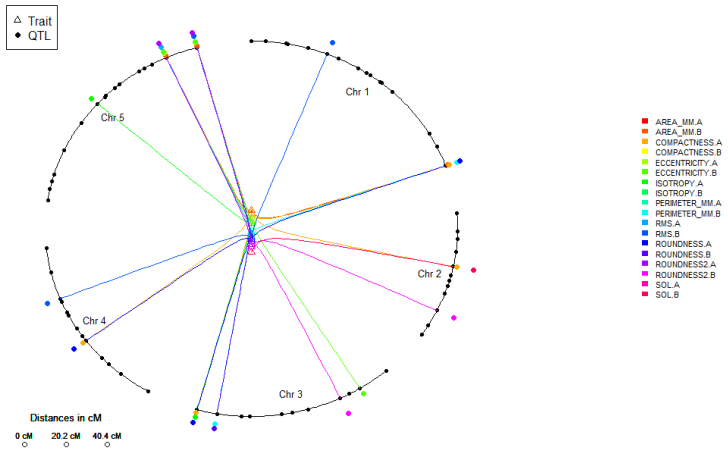
Figure 3.19: QTL profiles panel for Shape Descriptors and DAE(see text for graph description).)



(a) Raster image showing MQM results.



(b) Circle plot for all Descriptors and DAE pooled



(c) Circle plot for all Descriptors and DAE pooled

Figure 3.20: QTL profiles panel for Shape Descriptors and DAE (see text for graph description).

Descriptor	chr	pos..cM.	marker	LOD	5% Threshold	10% Threshold
ROUNDNESS_0	1	113.504	nga111	2.98	2.55	2.36
ROUNDNESS_1	1	124.009	MS_At1_29.6	2.74		2.60
COMPACTNESS_0	1	124.009	MS_At1_29.6	2.67		2.40
ROUNDNESS_2	1	124.009	MS_At1_29.6	2.50		2.48
ROUNDNESS_0	1	124.009	MS_At1_29.6	3.99	2.55	2.36
COMPACTNESS_2	2	10.932	MS_At2_2.4	3.40	2.85	2.31
COMPACTNESS_3	2	10.932	MS_At2_2.4	2.99	2.58	2.46
COMPACTNESS_1	2	10.932	MS_At2_2.4	2.55		2.41
COMPACTNESS_3	2	14.853	MS_At2_5.3	2.47		2.46
COMPACTNESS_2	2	14.853	MS_At2_5.3	2.50		2.31
COMPACTNESS_3	2	21.3	MS_At2_9.3	6.29	2.58	2.46
COMPACTNESS_1	2	21.3	MS_At2_9.3	4.64	2.94	2.41
COMPACTNESS_2	2	21.3	MS_At2_9.3	7.26	2.85	2.31
ROUNDNESS_2	2	21.3	MS_At2_9.3	2.98		2.48
COMPACTNESS_0	2	21.3	MS_At2_9.3	3.45	2.92	2.40
COMPACTNESS_4	2	21.3	MS_At2_9.3	3.73	2.84	2.57
COMPACTNESS_1	2	29.854	nga1126	6.14	2.94	2.41
COMPACTNESS_0	2	29.854	nga1126	4.84	2.92	2.40
COMPACTNESS_4	2	29.854	nga1126	2.94	2.84	2.57
COMPACTNESS_2	2	29.854	nga1126	7.47	2.85	2.31
COMPACTNESS_3	2	29.854	nga1126	5.86	2.58	2.46
COMPACTNESS_1	2	34.884	MS_At2_12.4	6.07	2.94	2.41
COMPACTNESS_2	2	34.884	MS_At2_12.4	9.20	2.85	2.31
COMPACTNESS_0	2	34.884	MS_At2_12.4	4.78	2.92	2.40
COMPACTNESS_3	2	34.884	MS_At2_12.4	7.73	2.58	2.46
ROUNDNESS_2	2	34.884	MS_At2_12.4	3.41	3.22	2.48
COMPACTNESS_4	2	34.884	MS_At2_12.4	4.95	2.84	2.57
COMPACTNESS_2	2	37.374	t32f646516	7.05	2.85	2.31
COMPACTNESS_3	2	37.374	t32f646516	5.81	2.58	2.46
COMPACTNESS_0	2	37.374	t32f646516	3.80	2.92	2.40

Table 3.5: Set of Markers over permutation threshold at 5% and 10% significancy level - Shape Descriptor by DAE (ShapeD_X) phenotypes

Descriptor	chr	pos..cM.	marker	LOD	5% Threshold	10% Threshold
ROUNDNESS_2	2	37.374	t32f646516	2.74		2.48
COMPACTNESS_1	2	37.374	t32f646516	4.60	2.94	2.41
COMPACTNESS_4	2	37.374	t32f646516	3.58	2.84	2.57
COMPACTNESS_1	2	40.474	MS_At2_14.9	2.84		2.41
COMPACTNESS_0	2	40.474	MS_At2_14.9	2.57		2.40
COMPACTNESS_2	2	40.474	MS_At2_14.9	4.44	2.85	2.31
COMPACTNESS_3	2	40.474	MS_At2_14.9	3.52	2.58	2.46
COMPACTNESS_3	2	44.08	t6a23ind10-10	2.58	2.58	2.46
COMPACTNESS_2	2	44.08	t6a23ind10-10	2.98	2.85	2.31
COMPACTNESS_3	2	54.615	athbio2b	2.48		2.46
ROUNDNESS2_4	2	54.615	athbio2b	2.50		2.35
ROUNDNESS2_4	3	16.768	nga162	2.88	2.75	2.35
ROUNDNESS2_4	3	22.457	MS_At3_6.5	3.49	2.75	2.35
ROUNDNESS2_4	3	27.007	msd2129380	3.31	2.75	2.35
ROUNDNESS_3	3	44.753	mzn14ind29-29	3.37	2.68	2.41
ISOTROPY_0	3	44.753	mzn14ind29-29	2.47	2.31	1.96
ROUNDNESS_3	3	52.389	k11j14ind16-16	3.31	2.68	2.41
ROUNDNESS_4	3	52.389	k11j14ind16-16	3.50	2.68	2.51
ROUNDNESS_3	3	57.967	MS_At3_16.0_a	3.39	2.68	2.41
ROUNDNESS_4	3	57.967	MS_At3_16.0_a	3.75	2.68	2.51
ROUNDNESS_4	3	73.811	MS_At3_18.2	4.38	2.68	2.51
COMPACTNESS_0	3	73.811	MS_At3_18.2	2.94	2.92	2.40
COMPACTNESS_2	3	73.811	MS_At3_18.2	2.84		2.31
ROUNDNESS_2	3	73.811	MS_At3_18.2	3.45	3.22	2.48
ISOTROPY_0	3	77.507	t16k521877	2.25		1.96
COMPACTNESS_2	3	77.507	t16k521877	5.75	2.85	2.31
COMPACTNESS_1	3	77.507	t16k521877	4.05	2.94	2.41
ISOTROPY_1	3	77.507	t16k521877	2.45		2.30
ROUNDNESS_4	3	77.507	t16k521877	5.46	2.68	2.51
ROUNDNESS_3	3	77.507	t16k521877	2.96	2.68	2.41
COMPACTNESS_4	3	77.507	t16k521877	3.11	2.84	2.57

Table 3.5: Set of Markers over permutation threshold at 5% and 10% significancy level - Shape Descriptor by DAE (ShapeD_X) phenotypes

Descriptor	chr	pos..cM.	marker	LOD	5% Threshold	10% Threshold
COMPACTNESS_3	3	77.507	t16k521877	3.81	2.58	2.46
COMPACTNESS_0	3	77.507	t16k521877	4.40	2.92	2.40
ROUNDNESS_2	3	77.507	t16k521877	5.69	3.22	2.48
ROUNDNESS_1	3	77.507	t16k521877	3.47	3.13	2.60
ROUNDNESS_2	3	89.533	athcdc2bg	10.37	3.22	2.48
ISOTROPY_1	3	89.533	athcdc2bg	4.65	2.91	2.30
ROUNDNESS_0	3	89.533	athcdc2bg	2.67	2.55	2.36
ROUNDNESS_1	3	89.533	athcdc2bg	5.96	3.13	2.60
COMPACTNESS_2	3	89.533	athcdc2bg	11.66	2.85	2.31
ISOTROPY_0	3	89.533	athcdc2bg	3.30	2.31	1.96
COMPACTNESS_1	3	89.533	athcdc2bg	8.45	2.94	2.41
COMPACTNESS_4	3	89.533	athcdc2bg	7.37	2.84	2.57
COMPACTNESS_0	3	89.533	athcdc2bg	6.33	2.92	2.40
ROUNDNESS2_2	3	89.533	athcdc2bg	2.97	2.57	2.36
ROUNDNESS_4	3	89.533	athcdc2bg	7.91	2.68	2.51
ISOTROPY_3	3	89.533	athcdc2bg	3.31	2.51	2.29
ROUNDNESS_3	3	89.533	athcdc2bg	6.44	2.68	2.41
COMPACTNESS_3	3	89.533	athcdc2bg	9.24	2.58	2.46
ISOTROPY_2	3	89.533	athcdc2bg	2.64	2.63	2.40
PERIMETER_MM_2	3	99.938	nga6	2.48		2.17
COMPACTNESS_0	3	99.938	nga6	8.24	2.92	2.40
COMPACTNESS_3	3	99.938	nga6	10.23	2.58	2.46
PERIMETER_MM_4	3	99.938	nga6	2.94	2.89	2.49
ROUNDNESS_1	3	99.938	nga6	7.41	3.13	2.60
ROUNDNESS_4	3	99.938	nga6	5.89	2.68	2.51
ISOTROPY_0	3	99.938	nga6	2.60	2.31	1.96
COMPACTNESS_1	3	99.938	nga6	10.14	2.94	2.41
ROUNDNESS_0	3	99.938	nga6	4.03	2.55	2.36
ISOTROPY_1	3	99.938	nga6	4.23	2.91	2.30
ROUNDNESS_3	3	99.938	nga6	6.08	2.68	2.41
ROUNDNESS_2	3	99.938	nga6	12.58	3.22	2.48

Table 3.5: Set of Markers over permutation threshold at 5% and 10% significancy level - Shape Descriptor by DAE (ShapeD_X) phenotypes

Descriptor	chr	pos..cM.	marker	LOD	5% Threshold	10% Threshold
COMPACTNESS_4	3	99.938	nga6	8.28	2.84	2.57
ISOTROPY_3	3	99.938	nga6	3.72	2.51	2.29
COMPACTNESS_2	3	99.938	nga6	13.63	2.85	2.31
ISOTROPY_2	3	99.938	nga6	3.13	2.63	2.40
ROUNDNESS_0	4	29.052	det1.2	2.79	2.55	2.36
ROUNDNESS_4	4	29.052	det1.2	3.40	2.68	2.51
ROUNDNESS_0	4	29.661	f28m11ind22-22	2.94	2.55	2.36
ROUNDNESS_4	4	29.661	f28m11ind22-22	3.67	2.68	2.51
PERIMETER_MM_2	4	41.003	ciw6	3.16	2.60	2.17
COMPACTNESS_1	4	41.003	ciw6	2.48		2.41
PERIMETER_MM_4	4	41.003	ciw6	3.61	2.89	2.49
COMPACTNESS_3	4	41.003	ciw6	2.83	2.58	2.46
ROUNDNESS_4	4	41.003	ciw6	5.15	2.68	2.51
ROUNDNESS_1	4	41.003	ciw6	6.56	3.13	2.60
ROUNDNESS_0	4	41.003	ciw6	6.83	2.55	2.36
PERIMETER_MM_3	4	41.003	ciw6	2.90	2.87	2.43
COMPACTNESS_0	4	41.003	ciw6	4.85	2.92	2.40
COMPACTNESS_2	4	41.003	ciw6	2.67		2.31
ROUNDNESS_2	4	41.003	ciw6	4.89	3.22	2.48
ROUNDNESS_3	4	41.003	ciw6	3.51	2.68	2.41
ROUNDNESS_0	4	44.925	MS_At4_8.3	4.87	2.55	2.36
COMPACTNESS_3	4	44.925	MS_At4_8.3	2.99	2.58	2.46
ROUNDNESS_2	4	44.925	MS_At4_8.3	3.88	3.22	2.48
ROUNDNESS_1	4	44.925	MS_At4_8.3	5.18	3.13	2.60
COMPACTNESS_4	4	44.925	MS_At4_8.3	2.88	2.84	2.57
ROUNDNESS_4	4	44.925	MS_At4_8.3	4.66	2.68	2.51
ROUNDNESS_3	4	44.925	MS_At4_8.3	3.20	2.68	2.41
COMPACTNESS_0	4	44.925	MS_At4_8.3	3.79	2.92	2.40
ROUNDNESS_1	4	49.896	MS_At4_9.3	3.14	3.13	2.60
ROUNDNESS_0	4	49.896	MS_At4_9.3	3.34	2.55	2.36
ROUNDNESS_4	4	65.517	f22k18ind3-3	2.96	2.68	2.51

Table 3.5: Set of Markers over permutation threshold at 5% and 10% significancy level - Shape Descriptor by DAE (ShapeD_X) phenotypes

Descriptor	chr	pos..cM.	marker	LOD	5% Threshold	10% Threshold
COMPACTNESS_4	4	65.517	f22k18ind3-3	2.67		2.57
ROUNDNESS_2	5	100.049	mqj2ind8-8	4.64	3.22	2.48
COMPACTNESS_3	5	100.049	mqj2ind8-8	3.49	2.58	2.46
COMPACTNESS_2	5	100.049	mqj2ind8-8	5.30	2.85	2.31
COMPACTNESS_3	5	85.452	mql58836	3.29	2.58	2.46
COMPACTNESS_2	5	85.452	mql58836	3.96	2.85	2.31
COMPACTNESS_2	5	85.491	nga129	3.97	2.85	2.31
COMPACTNESS_3	5	85.491	nga129	3.30	2.58	2.46
COMPACTNESS_4	5	88.673	MS_At5_21.3	3.14	2.84	2.57
ROUNDNESS_2	5	88.673	MS_At5_21.3	3.31	3.22	2.48
COMPACTNESS_3	5	88.673	MS_At5_21.3	4.09	2.58	2.46
ROUNDNESS_3	5	88.673	MS_At5_21.3	2.45		2.41
COMPACTNESS_2	5	88.673	MS_At5_21.3	4.90	2.85	2.31
COMPACTNESS_3	5	92.904	jv65-66	4.94	2.58	2.46
COMPACTNESS_2	5	92.904	jv65-66	6.10	2.85	2.31
ROUNDNESS_2	5	92.904	jv65-66	4.66	3.22	2.48
COMPACTNESS_4	5	92.904	jv65-66	3.73	2.84	2.57
ROUNDNESS_3	5	92.904	jv65-66	2.51		2.41

Table 3.5: Set of Markers over permutation threshold at 5% and 10% significancy level - Shape Descriptor by DAE (ShapeDescriptor_DAE) phenotypes

Descriptor	chr	pos..cM.	marker	LOD	5% Threshold	10% Threshold
AREA_MM.B	1	113.504	nga111	3.02	2.75	2.44
ROUNDNESS.A	1	113.504	nga111	2.81	2.51	2.24
ROUNDNESS.A	1	124.009	MS_At1_29.6	3.77	2.51	2.24
AREA_MM.B	1	124.009	MS_At1_29.6	3.68	2.75	2.44
RMS.B	1	38.807	f9h16ind26-26	2.56	2.34	2.12
COMPACTNESS.A	2	21.3	MS_At2_9.3	2.85		2.38
COMPACTNESS.A	2	29.854	nga1126	4.87	2.90	2.38
SOL.B	2	29.854	nga1126	3.07	2.53	2.15
COMPACTNESS.A	2	34.884	MS_At2_12.4	4.58	2.90	2.38
SOL.B	2	34.884	MS_At2_12.4	4.43	2.53	2.15
SOL.B	2	37.374	t32f646516	3.95	2.53	2.15
COMPACTNESS.A	2	37.374	t32f646516	3.75	2.90	2.38
SOL.B	2	40.474	MS_At2_14.9	3.33	2.53	2.15
COMPACTNESS.A	2	40.474	MS_At2_14.9	2.71		2.38
SOL.B	2	44.08	t6a23ind10-10	3.00	2.53	2.15
ROUNDNESS2.B	3	27.007	msd2129380	2.44	2.42	2.27
COMPACTNESS.A	3	73.811	MS_At3_18.2	4.12	2.90	2.38
COMPACTNESS.A	3	77.507	t16k521877	6.26	2.90	2.38
ROUNDNESS.A	3	77.507	t16k521877	2.43		2.24
ISOTROPY.A	3	77.507	t16k521877	3.14	2.78	2.25
ROUNDNESS.A	3	89.533	athcdc2bg	3.46	2.51	2.24
ISOTROPY.A	3	89.533	athcdc2bg	3.46	2.78	2.25
COMPACTNESS.A	3	89.533	athcdc2bg	8.93	2.90	2.38
ROUNDNESS.B	3	89.533	athcdc2bg	3.64	2.14	2.05
ISOTROPY.A	3	99.938	nga6	3.68	2.78	2.25
COMPACTNESS.A	3	99.938	nga6	9.81	2.90	2.38
ROUNDNESS.B	3	99.938	nga6	3.08	2.14	2.05
ROUNDNESS.A	3	99.938	nga6	4.75	2.51	2.24
COMPACTNESS.A	4	41.003	ciw6	4.14	2.90	2.38
ROUNDNESS.A	4	41.003	ciw6	6.33	2.51	2.24

Table 3.6: Set of Markers over permutation threshold at 5% and 10% significancy level - Intercept (A) and Slope (B) of a geometrical model

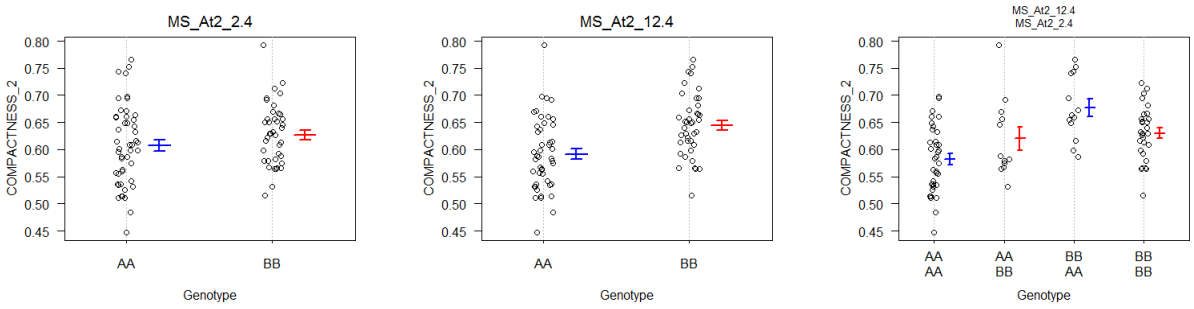
Descriptor	chr	pos..cM.	marker	LOD	5% Threshold	10% Threshold
COMPACTNESS.A	4	44.925	MS_At4_8.3	2.78		2.38
ROUNDNESS.A	4	44.925	MS_At4_8.3	4.66	2.51	2.24
ROUNDNESS.A	4	49.896	MS_At4_9.3	2.99	2.51	2.24
ROUNDNESS.B	5	88.673	MS_At5_21.3	2.19	2.14	2.05

Table 3.6: Set of Markers over permutation threshold at 5% and 10% significancy level - Intercept (A) and Slope (B) of a geometrical model

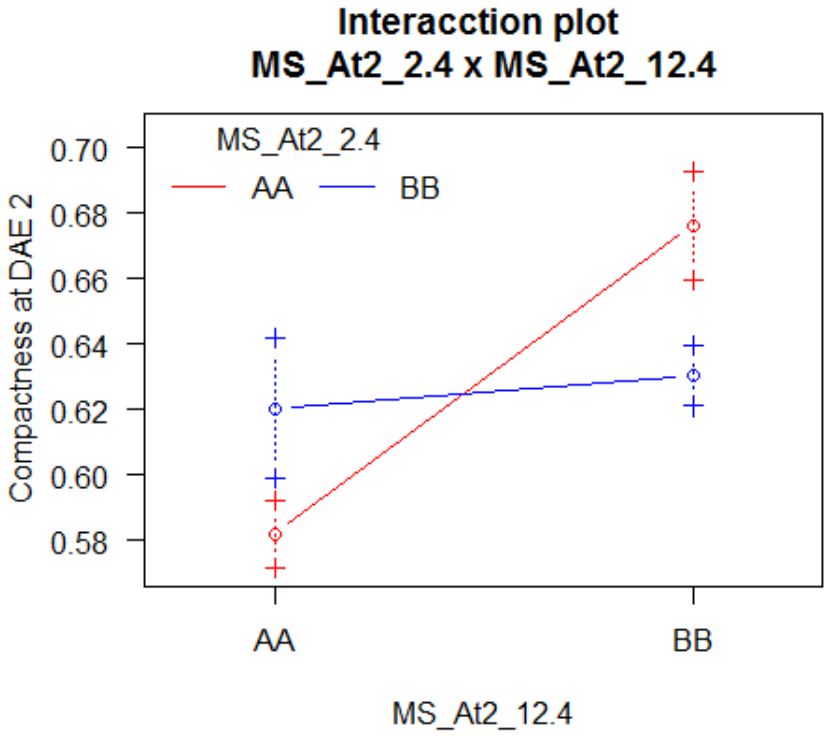
with 5% threshold of 2.53). The QTL at the end of chromosome 3 has a very high LOD for Compactness Intercept (LOD = 9.81 with 5% threshold at 2.90), but also Isotropy Intercept and Roundness Intercept and Slope. Some other possible QTLs appear by this approach. A possible QTL at the middle of the chromosome one is only supported by the slope of the Rotational Mass Symmetry (LOD = 2.56 over 5% threshold of 2.34). At the end of chromosome 1, Roundness Intercept and Area Slope (equivalent to growth rate) indicate a possible QTL with LOD scores over 3 and threshold around 2.50. A peak at the middle of chromosome 4 is significant for Roundness and Compactness Intercepts, with maximum LODs for Roundness Intercept at 6.33 and 4.66 with 5% threshold of 2.51. Finally, a possible peak at the end of chromosome 5 seems to be associated to Roundness Slope (LOD = 2.19 with 5% threshold of 2.14).

Selected examples have been chosen to further explore QTL results. The shape descriptor compactness at DAE 2 had LODs' over the 5% nominal threshold at markers situated at 10.9 and 34.9 cM. These markers, MS_At2_2.4 (marker A for short) and MS_At2_12.4 (marker B for short), had LODs of LOD(A) = 3.40 and LOD(B) = 9.20 ,over a 5% threshold of 2.85. The distribution of phenotypes at marker A (figure 3.21) indicates that genotype AA have a distribution that ranges from 0.45 to ∼0.75 with mean ∼0.61. Genotype BB ranges from 0.51 to 0.79 with mean ∼0.63. For marker B, genotype AA ranges from 0.45 to 0.79 with mean 0.59, but 0.79 seems to be a leverage point, since the next larger point is at 0.70, genotype BB range from ∼0.51 to 0.75 with mean 0.65. The mean differences assuming normal distributions explaining LODs over the 5% threshold. A look into the conditional distributions and interaction plots, a peculiar behaviour is found. It seems that marker 2 has a positive interaction with marker 1 when the latter has genotype AA but almost no interaction when it has genotype BB. This may be a signal of epistasis with opposite effects or somehow conditional interaction (see Mackay et al. (2009) for a review on detecting epistasis).

For comparison purposes, a similar analysis can be done by using the previous marker B with the marker athubique (marker C for short). Marker C did not result as significantly associated to Compactness_2. No epistatic interaction can be deduced from the interaction plot (not tested for significance, figure 3.22), since genotype BB at marker C increase average values identically for all genotypes at marker B. However, an additive effect of marker C seems



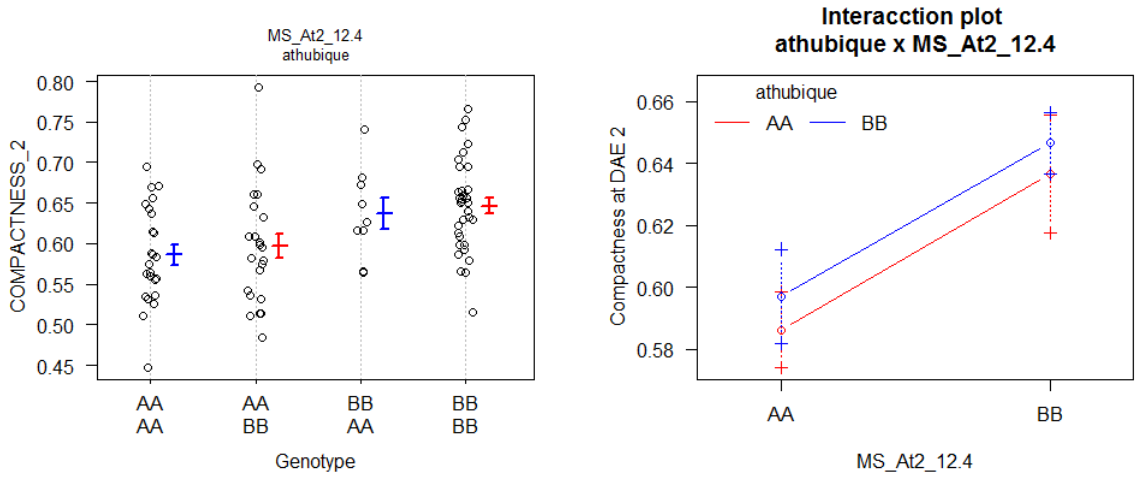
(a) (b) (c)



(d)

Figure 3.21: Marginal and conditional distribution of Compactness DAE 2 according to markers *MS_At2_2.4* and *MS_At2_12.4*
a) Marginal distribution for marker *MS_At2_2.4*
b) Marginal distribution for marker *MS_At2_12.4*
c) Conditional distribution according to both markers
d) Interaction plot for the genotypic effects of *MS_At2_2.4* and *MS_At2_2.4* on Compactness DAE 2.
The Interaction plots shows that marker 2.4 affects positively only to one genotype on marker 12.4 but not to the other.

to happen (not significant at 10% as extracted from LOD scores) since values for genotypes AA are lower than for BB.



(a) (b)

Figure 3.22: Marginal and conditional distribution of Compactness DAE 2 according to markers *athubique* and *MS_At2_12.4*
a) Conditional distribution according to both markers
b) Interaction plot for the genotypic effects of *athubique* and *MS_At2_2.4* on Compactness DAE 2.
The Interaction plots shows that marker *athubique* does not have epistatic interactions with *MS_At2_2.4* but both have additive effects on the phenotype .

3.4 Discussion

This experiment aimed to reveal Quantitative Trait Loci related with Arabidopsis thaliana rosette shape during its juvenile stage. Two natural ecotypes, Cape Verde Island (Cvi) and Argentat (Ag) - France, showed variation in rosette shape in the experiment described in chapter 2. A cross of these two accessions was bought for the experiment performed in this chapter. The population is a collection of RILs, produced by selfing for 8 generations after crossing them. Rosette shape of 8 replicate per RIL was measured using computer vision derived shape descriptors related with the distribution of pixels in top-view images taken daily.

The parentals, Cvi and Ag, develop their rosette with differences in rosette size, petiole length and leave shape. Visually, Cvi has shorter petioles and larger leaves than Ag, Cvi leaves are more elongated, while Ag are more rounded.

The F8 RILs had a distribution of shapes that range beyond the parentals, with multiple

combinations in phenotypes at lower level than the rosette, e.g petiole length, leaf shape and size or leave number. This phenomena is called transgressive segregation. Some RILs had very long petioles with small leaves, other small petioles and big leaves, or just small petioles and small leaves. It is remarkable the variation within plants in the pictures, where leaf shape change across new leaves appear and roll out. However, Shape Descriptors can hardly be associated to low level organ variation, accounting only for overall rosette shape. RILs transgressive segregation suggest that rosette shape is a polygenic trait whose combinations yield multiple outcomes due to epistasis, complementary genes, rare recessive alleles and overdominance (Rieseberg et al., 1999).

Thus, it is expected that rosette shape descriptors have a complex genetic architecture (Rieseberg et al., 2003), that require large populations and dense marker density for QTL mapping purposes. However, in our study, the experimental population is rather small, 89 lines, and the markers are sparse and in small number. Thus, it is expected than QTLs with major effects can be found rather than many small effects genes (Symonds et al., 2005; Soller and Beckmann, 1990).

The combination of several measurements of shapes and measuring days was expected to be helpful to avoid spurious false positive results and increase the opportunity of finding significant association. The modelling of QTLs for the five days separately and also as parameters of a geometrical model is a similar approach as taken by Moore et al. (2013). The idea is to evaluate the presence of similar QTLs from several, disparate , measurement of shape that covariates to some extent. Divergent timing in developmental transitions provide the advantage that QTLs may show different association degree at different time points. For that reason, QTL analysis is performed separately at each day. As commented by Moore et al. (2013), using function-valued trait, like fitting a geometrical model, provides a complementary view of the trait, removing noise and obtaining clearer results.

Several QTLs were consistently signalled as significant. The two most clear were at the middle of chromosome 2 and at the end of chromosome 3. Not so clear support is found for two other QTLs in the middle of chromosome 4 and at the end of chromosome 5. The one at the chromosome 4 emerge from roundness at several days, compactness at day 0, and roundness intercept, which correspond to the value at day 0, however is not reflected from the roundness

slope.It is not possible to narrow down these possible QTLs, since markers are around 7cM apart, and these intervals could contain many genes. The exploration of results yet may provide more interesting aspects on the the complex genetic architecture of shape descriptors Mackay et al. (2009). Although it has not been fully explored in this dataset, epistatic interactions seems to be widespread for morphological traits, suggesting complex regulatory networks on this kind of traits Mackay (2013). However, from a single experiment and single population, it is not possible to separate epistasis from variance due to allele frequency (Cheverud and Routman, 1995).

A comparison between the results in this experiment in the GWAS studing in chapter 2 is not straightforward. The genetic map in both populations is measured in different units (genetic (cM) vs physical (bp)), so no matching between regions can be performed. Broadly speaking, the most intense signal in this experiment, at the middle of chromosome 2 does not correspond to the potential QTL at the beginning of the chromosome 2 found in the GWAS. For the QTL at the end of chromosome 4, it may exist a similarity, since one of the QTLs found here and the QTL6 found in GWAS seem to be in similar region. However, the two potential QTLs that were found as not very significant, the one in the middle of chromosome 4 and the one at the end of chromosome 5 seems to overlap with QTL5 and QTL8, respectively, in the GWAS experiment. This may be indicative of the different genetic structure of both population. In the GWAS the genetic variation may have result in multiple rare variants controlling the same traits, in other words evolution may have found different solutions for the same problem. This population is less diverse, so less genes are expected to differ, and the QTL possibly at chromosome 4 and 5 may not be the most responsible for phenotypic variation in this population.

In order to study deeper these putative QTL, a population that allows for finer mapping is required. Making an introgression lines population between Cvi and Ag would be ideal to study separately the effect of these potential QTLs. In the literature no such lines have been found so that they could not be bought for successive experiments. However, the MAGIC population developed by Kover et al. (2009) from 19 natural accessions has shown phenotypic variation for rosette traits (Camargo et al., 2014) and will be use for fine mapping purposes in the next chapter.

Other approaches could be taken to analyse this population. For example, genotyping this

population with the same markers developed in the GWAS study would allow to compare maps in both experiments, but also to approximate finer to the QTLs position. Also (Gan et al., 2011) provide sequences for the 96 accessions studied by Atwell et al. (2010). Bioinformatic analysis of the sequence of Cvi and Ag could provide some insights on the genetic difference for candidate genes. Finally, the study of leaf descriptors, instead of whole rosette, would allow to study the correlation between traits such as petiole length and leaf size. The observations seems to indicate that in natural accessions petiole length and leaf size could correlate negatively, but the same my not happen in the RILs. If the traits are independent in the cross, this could indicate that both traits are genetically controlled by diffent pathway but in natural accessions the phenotypic correlation could be due to evolutionary forces. On the other hand, if the correlation persist in natural accessions and crosses may be indicative of common regulation and integration Searle (1961); Cheverud (1982); Wagner (1984).

Chapter 4

Association Mapping - Arabidopsis MAGIC Population

4.1 Introduction

Arabidopsis Multiparent Advanced Generation Intercross (MAGIC) population (Kover et al., 2009) has been phenotyped for rosette shape descriptor at juvenile stage for QTL fine-mapping purposes.

In chapters 2 and 3, two techniques for shape descriptors QTL mapping has been applied to a natural ecotypes population, GWAS, and a biparental cross population, Linkage Mapping. In both experiments a set of potential QTLs has been found but the mapping resolution was not enough to locate candidate genes in them. Association mapping on MAGIC populations offer the advantage of GWAS in natural populations, basically finer mapping than biparental crosses, without the drawback of false positives due to population structure. It is expected that Association Mapping on the MAGIC population would allow a deeper view on the genetic architecture of rosette shape.

In general, MAGIC populations are multi-parental crosses, typically 4,6 or 8, in a diallelic cross mating schema. The election of genotypically and phenotipically diverse parentals ensures enough genetic variation to explore the genetic architecture of complex traits. After crossing the parentals, several generations of selfing are achieved until ensure Recombinant Inbred Lines that are (nearly) genome-wide homozygous, so the population becomes genetically fixed. This "immortal" population can be phenotyped for several traits multiple times, since the genetic

composition of the lines does not change.

Kover et al. (2009) generated an Arabidopsis MAGIC population from 19 parentals, crossed for 4 generations and then selfed for 6 generations. The result is a population of 527 RILs whose genomes contain a mixture of the genome of 9.94 parentals in average. These RILs and the 19 parentals were genotyped for 1260 SNPs.

The association mapping method proposed for this MAGIC population has been implemented in a software package for R called *happy.hbrem*. The general strategy is as follows.

Each line is reconstructed as a mosaic of founder, i.e. parental, haplotypes. This means that for every SNP genotyped, the probability of Identity by Descent from every parental is calculated and with it the ancestral haplotypes. Thus, each SNP, instead of remain as biallelic polymorphysm, e.g A/G, has turned to have as many alleles as the number of founders, e.g SNP 1 at RIL 1 comes from Col-0 so has the Col-0 allele, and SNP 1 at RIL 125 coming from Ler-0 would have the Ler-0 allele. This is a midpoint strategy between using haplotype tagging and biallelic markers that allows to assign positive or negative effect in the trait to the founders' genome.

The association mapping strategy work through several stages. The first step is a standard genomic scan where each SNP is interrogated for statistical association between phenotypic and genetic variation, through a multiple linear regression model (equation 4.1). In this model, y is the phenotypic value, i is the individual (e.g RIL 1), L is the locus (e.g SNP 1), s is the founder haplotype (e.g Col-0), $P_{is}^{(L)}$ is the probability of the individual i having inherited the locus L from the founder s. $P_{is}^{(L)}$ is a weight for the phenotypic effect due to founder β_s.

$$y_i = \sum_s P_{is}^{(L)} \cdot \beta_s + e_i \tag{4.1}$$

A second step takes the result of the genomic scan to fit a multiple QTL model using the 80% of RILs. A resampling procedure, 500 repetitions, provides the measure of support for a QTL. Relevant details about the methods that goes beyond the interest of this thesis can be found at Kover et al. (2009); Valdar et al. (2009); Sen and Churchill (2001); Mott et al. (2000). According to Kover et al. (2009), the resolution of this method for association mapping in this population lies to around ∼6Mb corresponding to ∼300Mb, for traits with a 10% effect size.

In this chapter the MAGIC population grew into the PlantScreen phenotyping device, being

phenotyped daily for shape descriptors. Association mapping was performed for all descriptors and day combination and a Principal Components calculated on the descriptors.

4.2 Material and Methods.

4.2.1 MAGIC population

This population was generated from 19 parental accessions sequenced and analysed by Gan et al. (2011). The parental were intermated for four generations as described by Scarcelli et al. (2007) and inbred for other 6 generations. These RILs are considered homozygous for every marker, although some residual heterozygosity may remain. Hence, it is considered an 'immortal' population whose genotype does not change while reproduced by selfing. The whole set of RILs were genotyped for 1260 SNPs at a average distance of 100kb (Gan et al., 2011).

This population has been phenotyped for gene fine-mapping purposes in several papers. Kover et al. (2009) tested the population for time-to-event life-history traits such as Days to germination, Days to bolting, Days to flowering (under long and short days) and for morphological related traits such as Erecta and Glabrous phenotypes (two mutant phenotypes that affect the structure of the plant, Erecta, and the absence of trichomes, Glabrous). Other studies have focused in more ecological traits exploiting the genetic and phenotypic variation of the MAGIC population. As examples, Ehrenreich et al. (2009) and Banta et al. (2012) studied flowering phenology and its interaction with niche breadth, Gnan et al. (2014) analysed the genetic basis of seed size, number and its trade-off, Springate and Kover (2013) simulated climate warming and its effects on phenotypic plasticity (Springate et al., 2011).

4.2.2 Experimental Set up

The MAGIC RILs seedlings were grown initially at Aberyswyth University botanic gardens and later moved onto Natoinal Plant Phenomics Centers' PlantScreen phenotyping device. RILs were planted in three staggered, partially overlapping, assays (see figure 4.1). Each assay contained 3 replicates of, respectively, 161, 164 and 160 RILs, with no common RILs between them. Together with those RILs, three replicates of three natural accessions, Col-0, Ct-1 and Sf-2, were planted in the three assays (see figure 4.2). The role of these common accessions is

to serve as check of between assay similarity of conditions as can be seen at figure 4.4. As a note, only one replicate of Ct-1 germinate in the second and the third assay resulting in less replication than originally intended; One replicate of Sf-2 is missing also from assay 3.

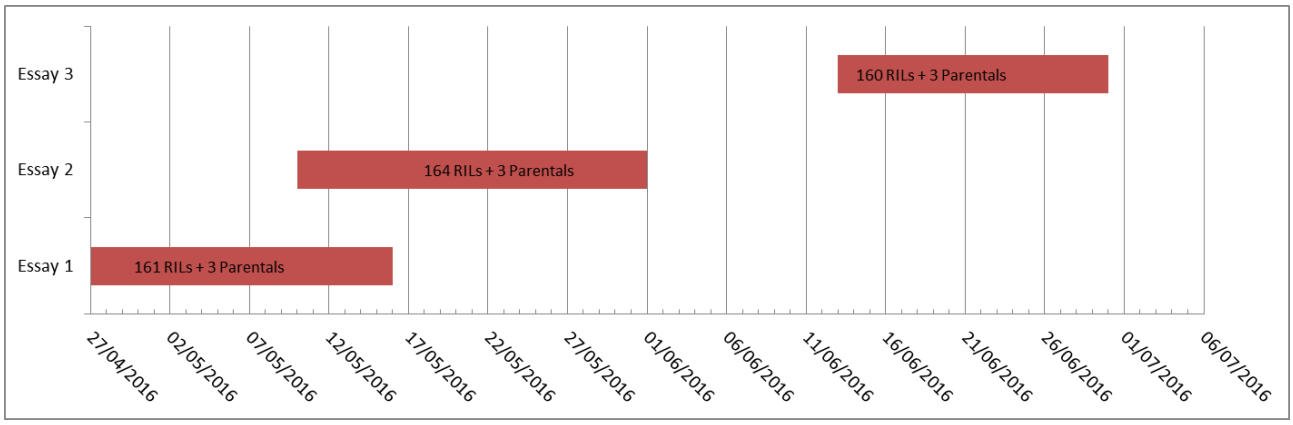


Figure 4.1: Gantt Chart showing Assay Temporal Schema – Starting date of every bar was the day that plants were placed into the PlantScreen Device

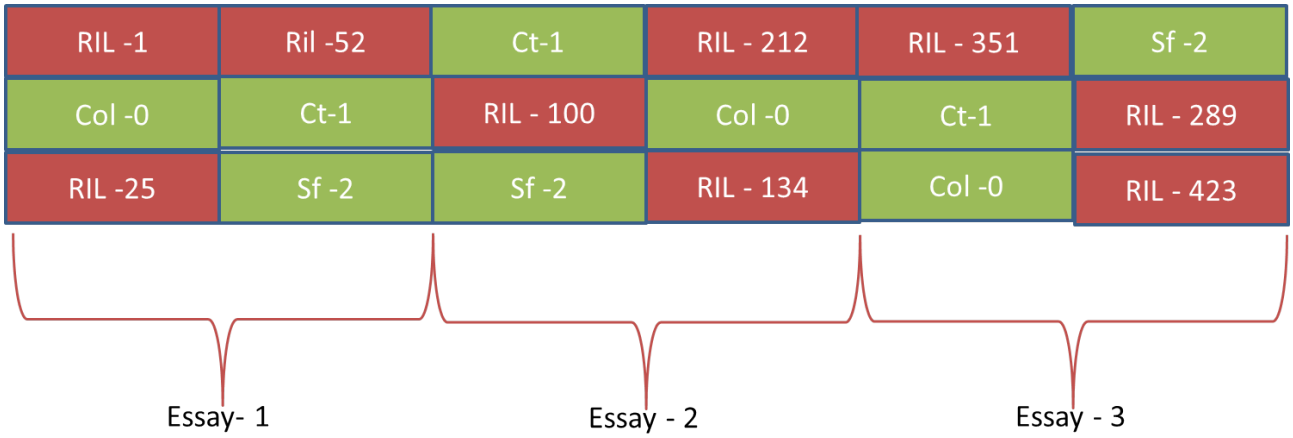


Figure 4.2: Schematic representation of RILs distribution in the experiment. Each "assay" contains $\sim$ 160 RILS. In a assay 3 replicates of each RIL plus 3 replicates of Ecotypes Col-0, Ct-1 and Sf2 are growing together in a randomized position. There is no common RILs between assays

Seeds were kept in vernalization for 28 days at 4°C in a dark room.After germination, seedlings were transferred to a glasshouse for 3 days at 14 hours day length and temperature range of 18°C, during day and 15°C, during night. Afterwards, individual seedlings were transplanted into circular 6 cm pots 50% filled with vermiculite at bottom (to restrict plant growth by keeping nutrient levels low) and top-filled with Levington F1/ 20% grit/sand compost. The seedlings were kept in the glasshouse after watering and covered with a transparent lid for 2 days for acclimation. Before going onto the conveyors, individuals were distributed in 5x4 pots per trays (figure 4.3) in three spatially randomised blocks. Each block contained a single

replicate of every RILs, so three blocks were equivalent of having three replicates of the assay. Three days later, all trays were placed into the PlantScreen (Photon Systems Instrument, Brno, Czech Republic) phenotyping device. Therefore, the three assays started 36 days after sowing. For easier interpretation, Days After Experiment started (DAE) counts as the number of days that plant were into the conveyor system. The translation from DAE to DAS would then be $DAS = 36 + DAE[days]$.

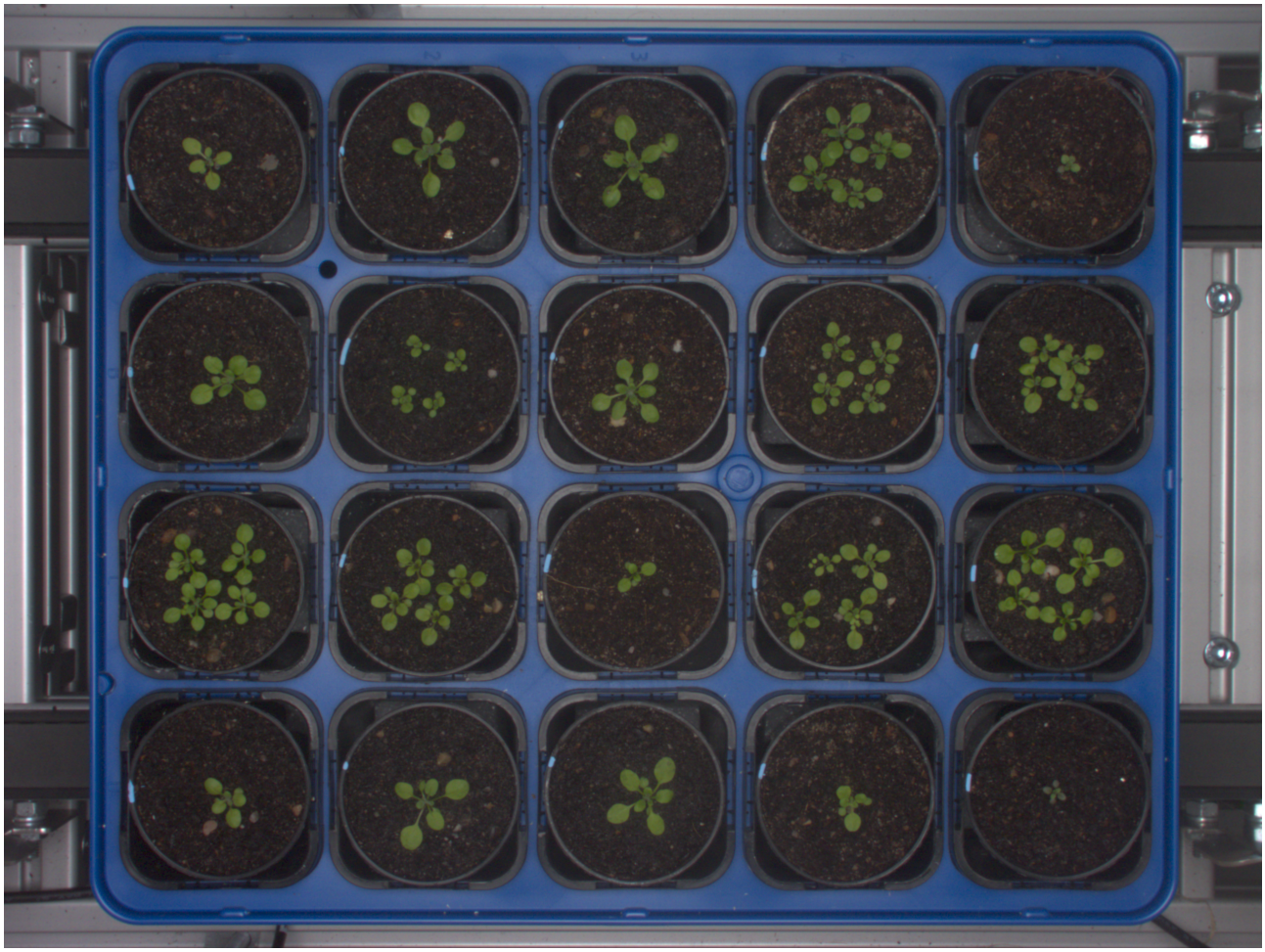


Figure 4.3: Example of MAGIC population experiment 5x4 tray(Pots with four plants belong to a parallel experiment).

The system watered pots daily to a pre-defined target weight of 75% of field capacity. The conditions were similar for the 3 assays performed. Assay 1 lasted for 20 days, while assay 2 lasted for 23 days and assay 3 for 18 days. The differences between assays duration was due to differential flowering time of individuals. Plants were maintained on the phenotyping device until all of them were flowering and their size were too large to being handled by the system. However, the analyses are performed for the first 10 days, due of some plants started flowering around that day, and algae started growing on the soil. Both, flowers and algae reduced the

quality of image segmentation dramatically.

4.2.3 Computational methods

Image processing and Shape Descriptors

The image processing and shape descriptors calculations are automatically performed by PlantScreen internal software. The explanation of the pipeline and shape descriptors calculated is provided in chapter 3. Shape descriptors were transformed to Principal Components in order to get variables for "pure" shape, independent of size and uncorrelated with any other. As it was observed in chapters 2 and 3, descriptors correlate due to shared information of rosette shape. The idea is to conflate the important shape components into several orthogonal variables.

Data Management

All data management has been performed using R programming language and a set of packages associated to it. Data management and graphs were made with extensively use of Hadley Wickham's set of packages, *plyr*, *tidyr*, *dplyr* and *ggplot2*. The function *princomp* was chosen for Principal component analysis. The function *lme* from package *lmer* was the election for the general lineal mixed models needed to calculate the Analysis of Variance using genotypes as random models. The function *varcomp* from the package *ape* aided to extract the components of variance for broad heritability calculations.

Broad-sense heritability

Broad heritability was calculated by a linear mixed model with the *lme* function (package *nlme*) in R. The model $ShapeDescriptor = W \cdot \sigma_{Assay} + Z \cdot \sigma_{RIL} + \sigma_{error}$ was used to separate the genetic variance, due to RILS, assay and the phenotypic errors. RILs identifier, e.g. RIL-1, serve as random factor accounting for genotype variation, noted as Z. Assay, noted as W is a fixed effect accounting for variation between assays. The environmental error is the the remainder error $\sigma_{Environment} = \sigma_{error}$. Broad heritability was calculated as $\frac{\sigma_{RIL}}{\sigma_{RIL}+\sigma_{Assay}+\sigma_{error}}$ using the function *varcomp* (package *ape*) in R.

Broad heritability provides the amount of phenotypic variability explained by the genotype variation. Low heritabilites, for example under 40%, would indicate that within genotype

variation is high, thus genotypes averages may be close but with high variability. This would mean that a rosettes of a genotype would be similar to a rosette of some other genotype rather to other rosettes of the same phenotype. This would be a falsation of our observation of similarity between ecotypes (see Introduction chapter). In addition, Association Mapping of traits with low heritability cannot be trust, since the mapping is based in the phenotypic differences between individuals with different genetic background. It is possible to quantify the heritability of a trait based in single or multiple genetic loci variation (Falconer and Mackay, 1996), although this approach is not taken here.

Association Mapping.

RIL-averaged traits were used as input for association mapping with the R package *happy.hbrem*. This package was specifically designed for MAGIC association mapping and originally used with the Arabidopsis MAGIC population(Kover et al., 2009).

The function *prepare.database()* collect TAIR9-based SNPs physical map and RILs genotypes to reconstruct the parental haplotype mosaic of RILs chromosomes. The details of the method are beyond the interest of this chapter. It uses a dynamic programming algorithm, calling a Hidden Markov Model, to calculate the probability of a SNP being descent from a founder according to upstream markers (Kover et al., 2009). Kover et al. (2009) assure that haplotype reconstruction ascertain the founder of origin "with high probabilities", with the exception of chromosome borders, centromeres and recombination breakpoints. The result of this step is saved into a set of files and R scripts available for subsequent QTL mapping analysis.

The function *scan.phenotypes* read files containing a data matrix with MAGIC RIL per row and phenotypes on columns. First, the *happy* sub-function performs a genome scan for evidence of a QTL associated with the phenotype. The next step, a second method named *Hierachical Bayesian modelling*, or *hbrem*, is called on the results of the genomic scan. The function fits a model of phenotypic values on the genotypic value of every SNP. It is done using a random effect model, calculating the proportion of variance explained by each founder out of the total phenotypic variance.

The last step in the procedure is a 500 times resampling and multiple QTL model fitting. This serves to measure the support for QTLs. The result is a genome-wide value that in case

to match a genome-wide threshold of 5.8% may indicate the presence of a QTL on an interval (Kover et al., 2009). Finally, QTL location is then defined as the marker with largest -logP-value (from the hbrem step), and the interval where other SNPs reach a large enough genome-wide P-value (from the resampling procedure).

The function returns a list of possible QTLs, with the significance as logarithm of p-values, the genome-wide value and the position of the peak marker, also the calculated interval region with genome-wide value over the threshold (5.8%). In addition, the software saves the influence of the imputed parental-of-origin for every SNP marker that shows statistical association with the phenotype and boxplots of phenotypic distribution for each (parental) allele at each QTL. Finally a genome-wide Manhattan-like plot of SNPs p-values corrected by permutation test helps to discriminate potential QTLs.

QTL mapping algorithms were applied to our Shape Descriptors by DAE traits. First a model without covariables was used and afterwards a model with the markers ER_475 as covariate to compensate for peaks close to Erecta (ER) gene at chromosomes 2 .

Shape QTL *a posteriori* Analysis

Every Shape Descriptors - DAE combination results in either 0 or several genome intervals associated to them. The resulting intervals overlap in different descriptors and DAEs. The *union* of overlapping intervals in different traits provides a rough description of the broad genetic architecture of shape related traits and diminish the redundancy in the intervals.

Several stragegies has been followed to identify potential causal genes within broader loci.

Initially, every peak marker, i.e. signalled by association mapping as a significant hit, was searched in the TAIR9 database (`arabidopsis.org`) to locate which gene, if any, was underlying the marker. For each of those locus, the gene description was inspected to observe any known possible relationship between leaf, petiole or rosette and shape, size or developmental biology.

Secondly, whole set of markers within joint intervals were searched in the same TAIR9 database, retrieving all annotated genes within or overlapping such intervals. This was performed using the software suite *bedtools* in a Linux based High Performance Computer, and the TAIR9 annotated genome file.

Finally, all genes found, either under peak markers or within intervals, were confronted

against a public dataset of Arabidopsis rosette shape related genes named PhenoLeaf (Wilson-Sánchez et al., 2014). Gene Enrichment was calculated by means of the hipergeometrical test in R. This Gene Enrichment was applied separately to the genes under the peak markers, the genes within the joint intervals and those genes from PhenoLeaf that were found either under peak markers and within the joint intervals.

4.3 Results

4.3.1 Phenotypic Variation and correlation.

A first exploration of experimental results allows us to evaluate the homogeneity between the three assays. It is done by checking the phenotypic variation on parental accessions (figure 4.4) and the set of all RILs (figure 4.5).

Area and perimeter show an exponential growth throughout time as would be expected from a juvenile plant growing. However, the shape descriptors have a more variable pattern along time. Rosette area growth shows accession Ct-1 growing faster than Col-0 and Sf-2, being this two similar. However, looking at the perimeter, it is Col-0 accession which grows at a slower rate than the other two accessions, that keep growing similarly. This may indicate variation of shape captured by differences in the relationship area-perimeter for a rosette, which is actually exploited by roundness.

The other shape descriptors can be grouped according to their behaviour along time. Isotropy and roundness are decreasing through plants development. Isotropy for Col-0 do not decrease after day 5 as the other two accessions do. Roundness shows a similar pattern, except that is negatively decreasing from DAE 0 to 10. Sf-2 keeps a less round pattern than Col-0. Ct-1 seems to be more similar to Col-0 during the first 5 days, and closer to Sf-2 by the end of the experiments. Eccentricity, roundness2 and RMS have a similar pattern between them. Eccentricity and RMS decrease through time for Col-0 and Ct-1 but not for Sf-2. This may be the result of longer petioles in Sf-2 after day 5, when start increasing. The slope for Sf-2 accession in these descriptors separate it from the others two. Roundness2 show the inverted behaviour than RMS and eccentricity, that is, when RMS increases roundness2 decreases and vice-versa. Compactness seems to be constant for accessions Col-0 and Ct-1 but decreasing for

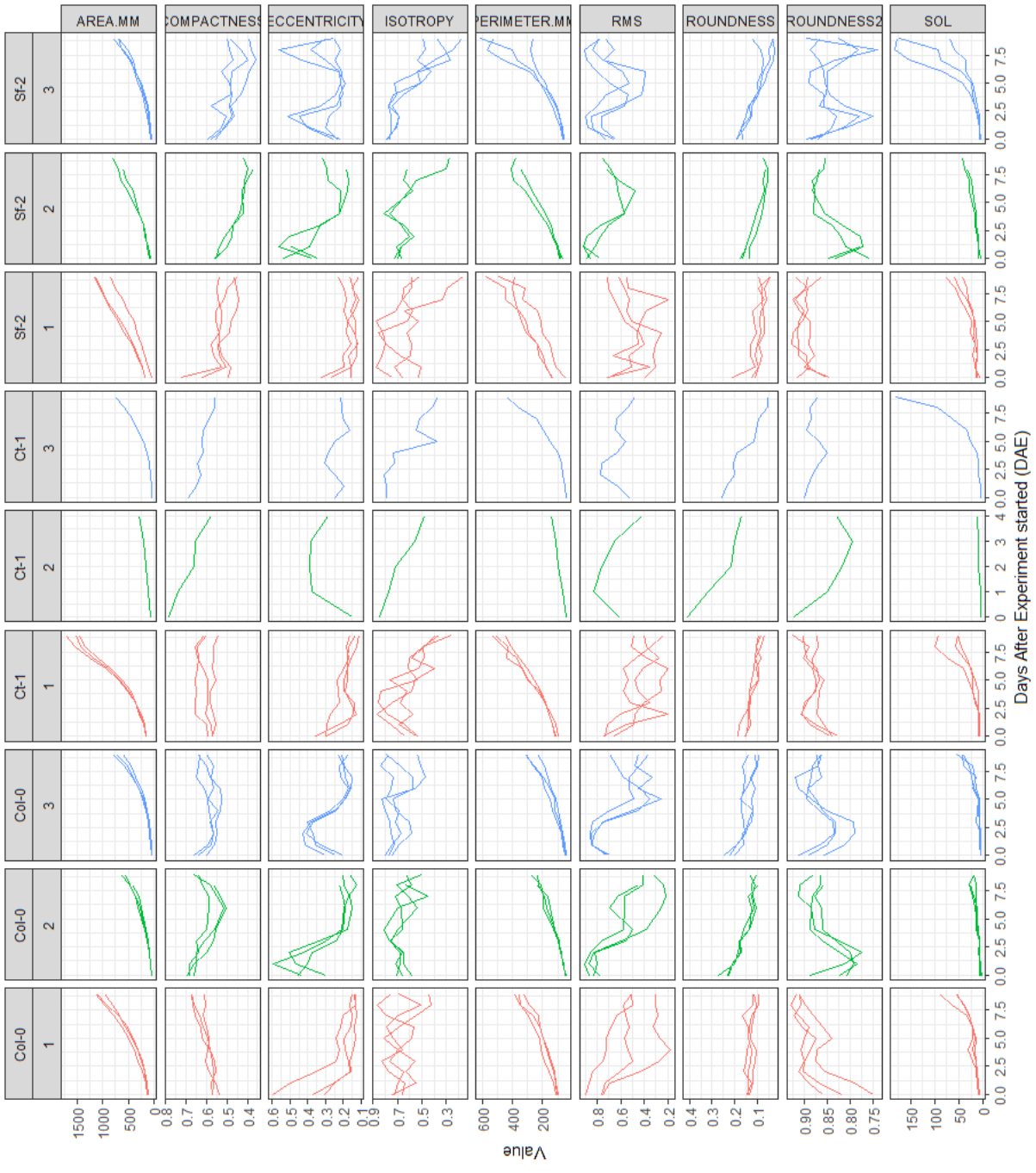


Figure 4.4: Shape Descriptors Trajectories through time – Parental Natural Accessions comparison between assays. Columns numbered 1,2 and 3 correspond to the equivalent assay. Colours correspond also to assay for easier comparison. Values are expresed in squared millimetres for area and millimetres for perimeter, the rest are arbitrary units

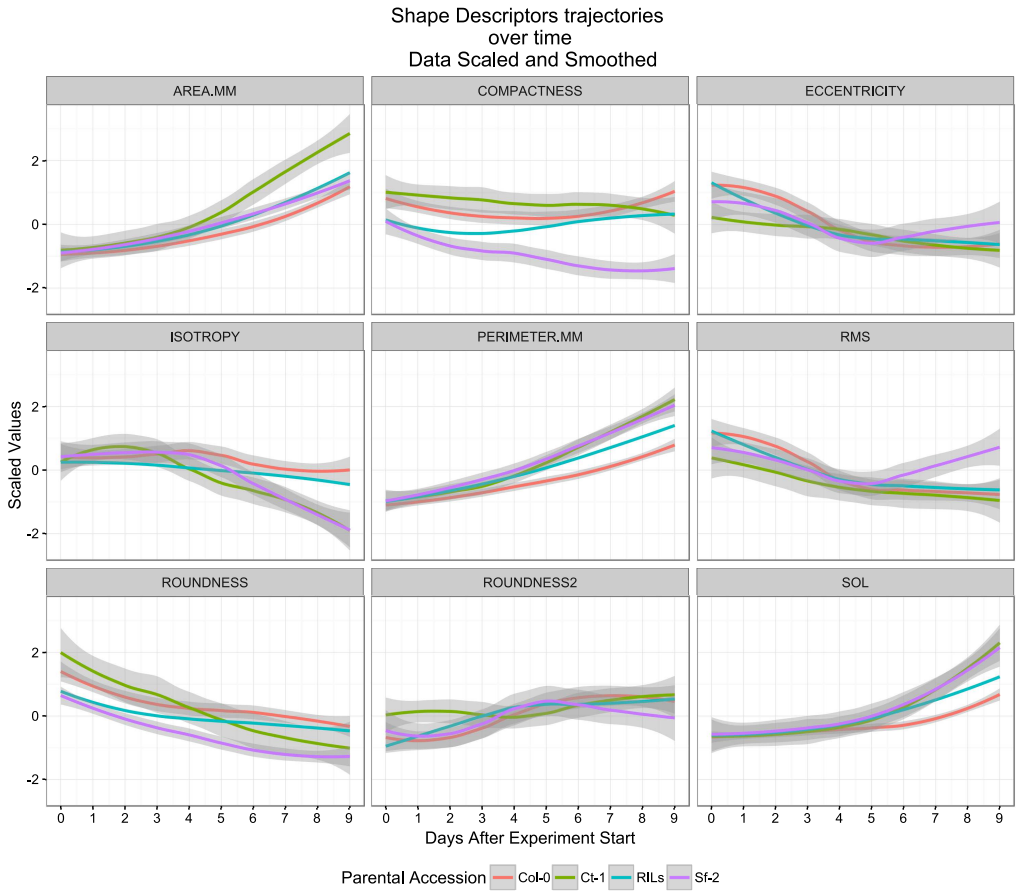


Figure 4.5: Shape Descriptors Trajectories through time –Parental Natural Accessions and RILs. Variables have been scaled ($\frac{value-\mu}{\sigma}$) for comparison purposes. Variance has been calculate by the loess procedure (smooth by local polynomial regression fitting)

Sf-2 suggesting that it discriminate between dense/loose rosette habit.Slender of leaves seems correlating with perimeter, growing faster for Ct-1 and Sf-2 than for Col-0.

For all the descriptors time trajectories, the pooled values for RILs are kept in between the parental strains (figure 4.5). This is expected since they have been pooled so number of plants is very large so variance is very small.

In summary, the three parental accessions show differences between the descriptors, with different descriptors able to group differently the time trajectories of them. Using the parental ecotypes as examples we can assume that these descriptors are valid to account for differences between rosette shapes, but a similar analysis for RILs would be impracticable due to the high amount of genotypes. Figure 4.4 suggest that the three parental accessions may have been growing faster in the assay 1 than in the other 2.

Correlation between descriptors

Pearson correlation between Shape descriptors confirms the similarities among their pattern through time (see figure 4.6). Isotropy, roundness and compactness are correlated positively, in a range between 0.49 to 0.7. Roundness2 and eccentricity are strongly correlated with $|corr| > 0.8$. Finally, area and perimeter are correlating with a value of 0.7. The correlation between Shape Descriptors and area or perimeter is over 0.5 for most of the descriptors. Only compactness and roundness does not correlate with neither area nor perimeter.

The correlation between shape descriptors and size-related variables, i.e. perimeter and area, support to perform a Principal Component analysis to separate the influence of size into shape.

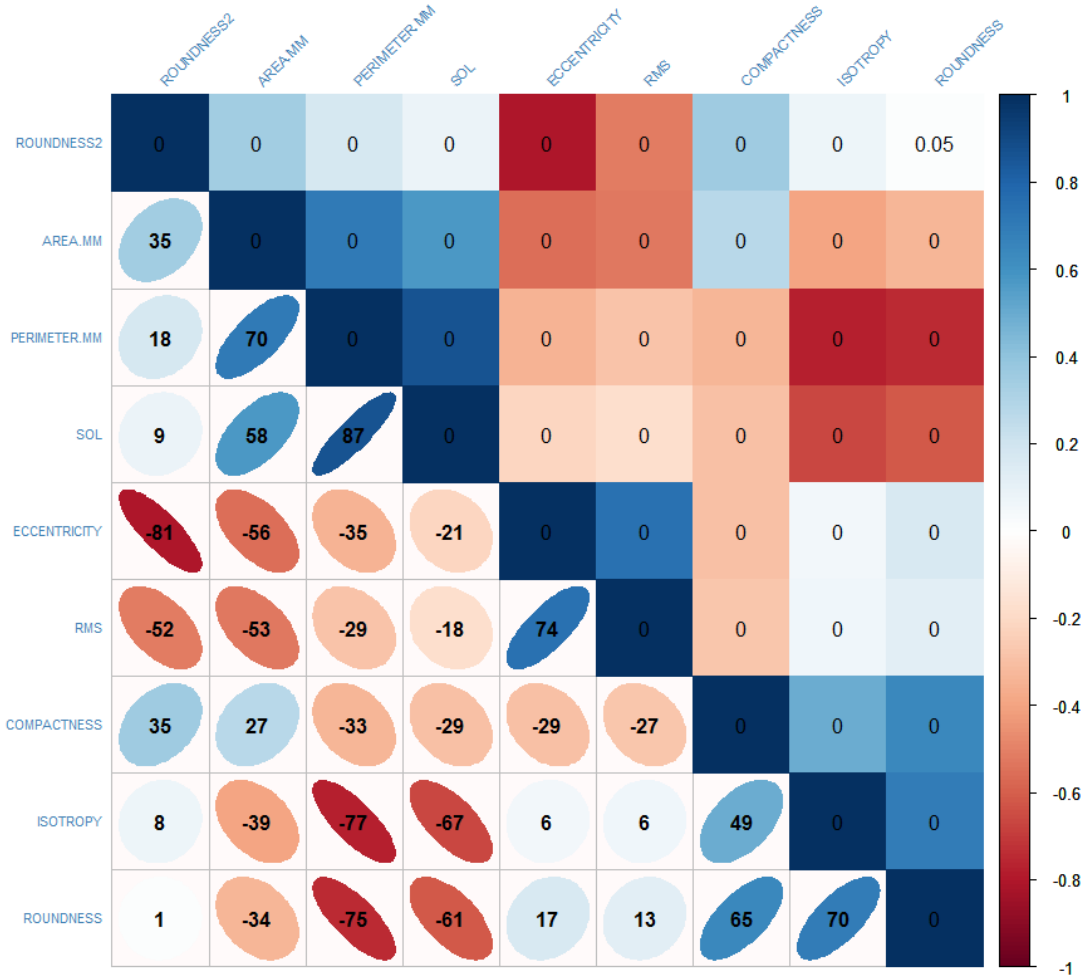


Figure 4.6: Correlation Plot of Shape Descriptor Values - Pearson correlation is indicated by color, number enclosed in the squares and ellipse eccentricity

Principal Component Analysis

Principal Component Analysis (PCA) is performed using the correlation matrix calculated over RILs indidual values, rather than averages, of the nine Shape Descriptors and pooling values from all days. Eigenvectors were calculated to obtain the Principal Component loadings and their associated eigenvalues. Eigenvalues provide the amount of variation explained. Figure 4.7a indicates that the first PC retain almost 50% of variance in shape descriptors, and the second rise up to around 75%.

Principal components coefficients (see figure 4.7b and Table 4.1) show that Principal Component 1, 3, 5, 6 and 9 are influenced by area. Principal Component 1 is also influenced by perimeter, so we can assume that PC1 retains most of the variation due to size, while PC2 contain most of the shape variation. The rest PCs, from 3 to 9, seem to have a more mixed combination of shape and size.

To help to visualize the effect of PCA transform on measurements, figure 4.7b shows PCA loadings on the PC1-PC2 space and figure2 4.7c and 4.7d illustrate the rotation of values in the axis PC1-PC2. Principal Component 1 retains 47% of variation by including area (0.42), perimeter (0.45) and SOL (0.37) as their main positive contribution, but also RMS (-0.36) and eccentricity (-.38) have large contribution on the negative side. Principal Component 2 retains 24% of variation (71% of cumulative variation) by its contribution of compactness (0.54), roundness (0.42) and isotropy (0.41), but again RMS (-0.22) and eccentricity (-0.29) plays a negative role in this component.

To visually clarify the Principal Component Analysis, figures 4.7c and 4.7d presents two versions of the same scatter plot of PC1 over PC2 for RILs rotated values of Shape Descriptors. Figure 4.7c is coloured by the original values of area and figure 4.7d by original values of compactness. Colour gradient in both images indicate that area increase from left to right in Principal Component one and compactness does it bottom up.

An example of segmented plant has been plotted at the bottom of figure 4.7c. Its PC values has been plotted along its development in the PC1 vs PC2 scatterplots with a black path. This example can be used to explore the interpretation of shape descriptors through time. Plant develops and growth over time, so the path moves from left to right side in PC1. At the same time, for the 4 first days, the trajectory indicates that the plant are heading towards more

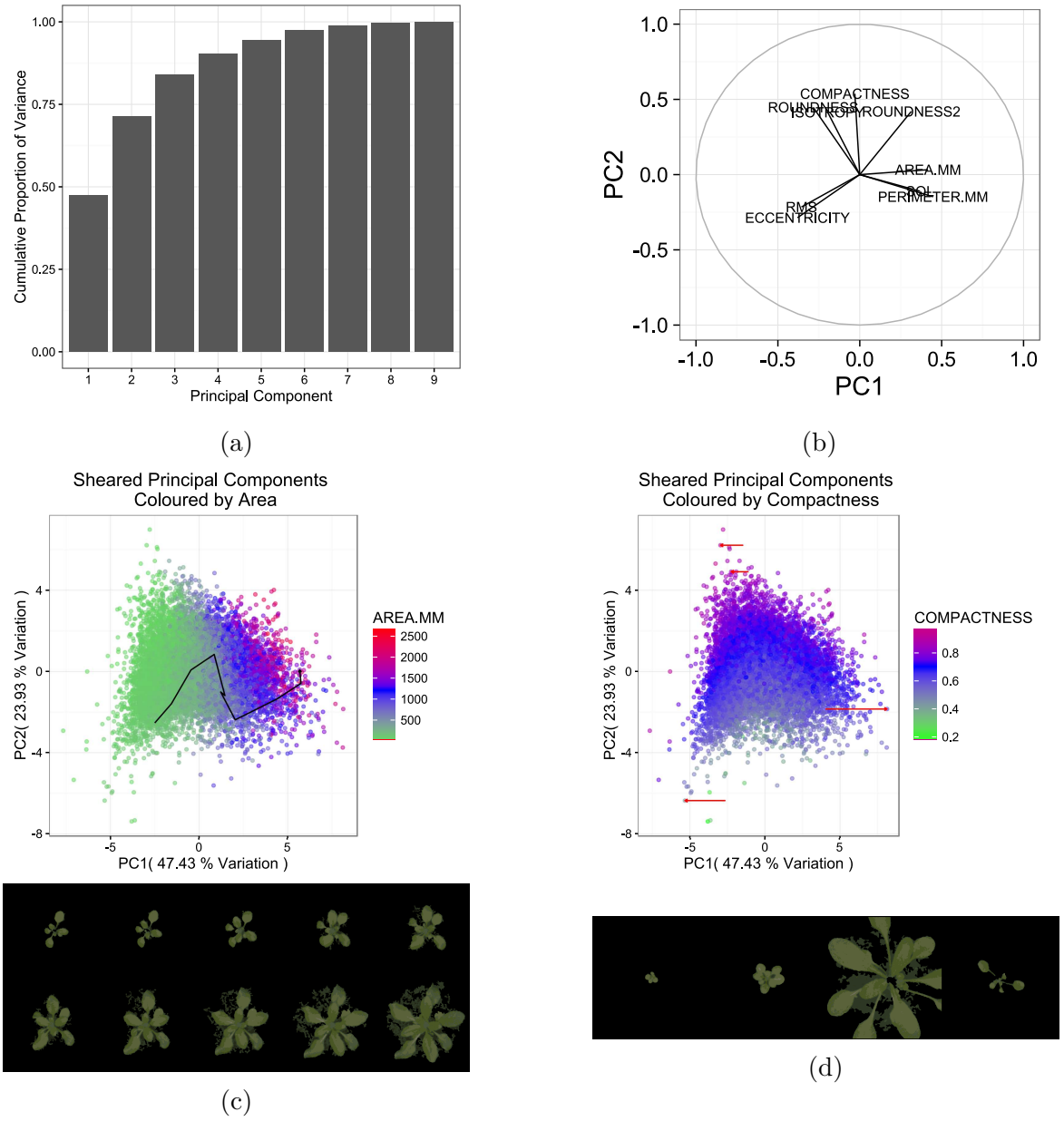


Figure 4.7: RILs Shape Descriptors Principal Component Analysis. a)top-left: Cumulative Proportion of Variance explained by Principal Components b)Top'Right: Loadings of variables into Principal Component 1 and 2 c) Bottom Left. Rotation of RILs Shape Descriptors to Principal Components 1 and 2. Color correspond to area. Black path, Trajectory of one example plant throughout days. Bottom, images corresponding to plant rosette in the example d)Bottom Right. Rotation of RILs Shape Descriptor to Principal Components 1 and 2. Color corresponds to compactness. Red Arrows correspond to four examples shown at the bottom of the graph. From left to right matches examples from top-down

positive values of PC2. This is due to some new leaves appearing, existing leaf blades are growing but petioles remain short. In this stage, the rosette has a more compact and round habit. After day 4, a sudden drop in PC2 and shorter movement in PC1 indicates that growth rate may have been reduced and the rosette becomes more sparse. A possible explanation is that plant has become more eccentric due to the "northest" leave expansion. For the last 2 days, the plant continues growing but the segmentation artefacts (algae growing on soil) make the phenotypic values on PC2 rise again toward more compact and round values.

Figure 4.7d has 4 examples of rosettes at the bottom. The four rosettes are signalled in the PC1-PC2 scatter plot by red arrows. The arrow at top of PC2 correspond to leftmost rosette picture, the second highest correspond to the second from left picture and so on. Small rosettes are located at the left side of PC1, corresponding pictures 1,2,4. Compact rosettes, pictures 1 and 2, are on the upper part of PC2. An eccentric rosette, e.g picture 4, is on the bottom part of the PC2. The biggest rosette, e.g picture 3, is at the rightmost side of PC1. With the examples and loadings plot, it has been shown that can be assumed that PC1 contain most of the variation in size (either area or perimeter), while PC2 contains mostly information about shape.

Table Descriptor	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
AREA.MM	0.42	0.03	0.31	-0.18	0.38	-0.47	0.16	-0.07	-0.55
COMPACTNESS	-0.03	0.54	0.53	0.06	0.09	0.07	-0.63	-0.1	0.12
ECCENTRICITY	-0.38	-0.29	0.36	-0.1	0.17	0.12	-0.04	0.75	-0.16
ISOTROPY	-0.2	0.41	-0.25	-0.83	0.14	0.1	0.12	0.03	0.04
PERIMETER.MM	0.45	-0.15	0.17	-0.18	0.13	-0.23	0.08	0.24	0.76
RMS	-0.36	-0.22	0.32	-0.28	-0.57	-0.51	-0.01	-0.24	0.05
ROUNDNESS	-0.28	0.45	0.3	0.31	0.06	-0.04	0.71	0	0.16
ROUNDNESS2	0.31	0.42	-0.19	0.08	-0.57	-0.2	-0.02	0.54	-0.17
SOL	0.37	-0.11	0.42	-0.24	-0.36	0.63	0.24	-0.12	-0.12

Table 4.1: PCA loadings calculated from Shape Descriptors correlation

Heritability

Broad heritability is generally defined as the amount of phenotypic variation explained by genotypic variation. In here, it helps to elucidate if the population shows enough divergence, at least some genotypes, to perform and validate the QTL mapping .

Table 4.2 shows the results for the broad heritability per Shape Descriptor and Day. The

Descriptor	0	1	2	3	4	5	6	7	8	9
AREA.MM	0.62	0.68	0.68	0.68	0.69	0.7	0.68	0.66	0.65	0.62
COMPACTNESS	0.69	0.7	0.64	0.62	0.57	0.52	0.49	0.49	0.52	0.56
ECCENTRICITY	0.39	0.48	0.52	0.47	0.37	0.37	0.18	0.27	0.28	0.28
ISOTROPY	0.16	0.23	0.22	0.23	0.26	0.25	0.23	0.22	0.17	0.21
PC1	0.46	0.6	0.66	0.62	0.6	0.57	0.48	0.47	0.4	0.36
PC2	0.61	0.58	0.52	0.53	0.52	0.5	0.42	0.41	0.4	0.47
PC3	0.43	0.57	0.6	0.49	0.36	0.32	0.27	0.22	0.2	0.19
PC4	0.3	0.21	0.14	0.13	0.13	0.13	0.15	0.15	0.08	0.06
PC5	0.31	0.24	0.16	0.13	0.17	0.18	0.22	0.19	0.15	0.21
PC6	0.19	0.18	0.12	0.16	0.23	0.26	0.21	0.24	0.24	0.24
PC7	0.36	0.4	0.34	0.29	0.31	0.31	0.29	0.32	0.3	0.23
PC8	0.31	0.32	0.24	0.22	0.14	0.09	0.1	0.1	0.1	0.11
PC9	0.42	0.4	0.36	0.37	0.3	0.19	0.23	0.31	0.33	0.41
PERIMETER.MM	0.67	0.74	0.75	0.72	0.72	0.69	0.62	0.58	0.46	0.47
RMS	0.39	0.46	0.43	0.32	0.21	0.17	0.1	0.11	0.17	0.18
ROUNDNESS	0.64	0.7	0.67	0.61	0.59	0.55	0.48	0.46	0.43	0.48
ROUNDNESS2	0.47	0.44	0.43	0.41	0.35	0.34	0.27	0.29	0.28	0.29
SOL	0.62	0.62	0.58	0.49	0.34	0.31	0.21	0.18	0.14	0.12

Table 4.2: Shape Descriptors By Day - Broad Heritability

descriptors area and compactness are well explained by genetic variation for the 10 days of the experiment, with values over 0.5. Perimeter keeps over this value only for the first 7 days, roundness for 5 days and slender of leaves only for three days. RMS, eccentricity and roundness2, which correlate between them, are only slightly heritable for the 3 first days and do not pass the threshold of 0.5. The first two Principal Components show heritabilities over 0.5 only for the 5 first days, but the rest of PC's only seems associated to genetic variation on sparse days.

4.3.2 Quantitative Trait Loci Mapping

The software *happy.hbrem* using a model without covariables found 227 SNPs with a significant association with Shape Descriptors. The raw results has been collected in the appendices B.1, that contain the peak markers and the intervals found significant by the software, and B.2, that contain the effect of each parental on the phenotype for every marker found significant. A summary with the number of potential QTLs is found at the table 4.3 and figure 4.8 shows the distribution of QTL markers by Trait and DAE along the genome.

Candidate QTLs are represented by a "peak SNP marker" and a chromosome segment. The segments contains markers with genome-wide p-value over a significance level established by

permutation test. Peak marker is the one having highest significance p-value in the segment. The union of overlapping segments allows gather redundant QTLs regions between Shape Descriptors and DAE into possible, more general, QTLs segments. The union of the 227 segments results in 43 intervals of size ranging from 207 to 30384507 base pairs. At the table B.1 the joint intervals are represented in black at the top of the figure.

As stated before, the effect size for each parental is estimated. The table B.2 contain the effect size for every parental per trait and peak marker. A brief descriptions of the most relevant intervals is provided, followed by an example of allelic distribution of ER_472 on PC2 on day 6.

The largest interval is located at chromosome 2 and might be argued that it is gathering several smaller intervals. In this region the marker ER_472 is found over the gene *ERECTA*, which is known to influence the plant architecture (see discussion). For that reason, Shape Descriptor data has been re-analysed to incorporate ER_475 as a covariate that could result in a separation of subintervals. Notice the selection of ER_475, nearby ER_472, to remove the influence of *ERECTA* in other markers, but allowing signal still happening in ER_472.

The QTL mapping using ER_475 as covariable found 217 significant QTLs (see table 4.4 and figure 4.9). Joining the segments as explained before, 41 intervals are found. The lost intervals is due to gathering 3 segments in the middle of chromosome 5 as single one. Remarkably, chromosome 2 big interval remain similar to the model without covariables, suggesting there is not important effect in using the gene ERECTA (located under ER_475) as covariable

The trait Principal Component 2 on day 6 had was associated to the peak ER_472 with -log p-value $\sim$10. The effect size for every founder allele is plotted in the figure 4.11. RILs with the ER_472 alleles imputed to the parentals Can, Hi and Ler have higher values of PC2 than all the others (tables 4.5 and 4.6). This example show that PC2 is affected by the shape of the rosette, since RIL-16 (Ler allele) and RIL-164 (Can allele) examples are very compact, as opposite to RIL-119 (Col allele) and RIL 522(Sf allele). The examples of RIL-317 (Hi) allele and RIL-507 (Bur allele) are in between those values. Also, Ler, whose completed name is Landsberg erecta, has a mutation on the gene ERECTA that has strong effect on leaves and inflorescence structure, which is consistent with the result obtained for the alleles at PC2_6.

Trait/DAE	0	1	2	3	4	5	6	7	8	9	Trait Total
AREA.MM	0	0	0	0	0	0	0	0	0	2	2
COMPACTNESS	11	4	2	5	2	3	1	8	5	3	44
ECCENTRICITY	0	0	0	4	0	0	1	0	0	0	5
ISOTROPY	0	3	0	1	4	1	1	1	1	1	5
PC1	0	0	0	1	0	0	1	3	3	4	12
PC2	3	2	2	2	3	1	1	5	4	8	31
PC3	1	2	1	8	1	0	0	0	0	1	14
PC4	0	0	0	0	3	2	0	0	0	0	5
PC5	2	0	0	1	0	0	0	0	0	0	3
PC6	0	0	0	1	0	0	0	1	4	0	6
PC7	1	1	0	0	0	0	0	0	0	0	2
PC8	0	0	2	0	0	0	2	0	0	0	4
PC9	1	2	1	1	1	0	0	0	0	0	6
PERIMETER.MM	0	0	2	3	2	1	2	2	5	4	21
ROUNDNESS	4	4	7	1	2	1	4	3	6	8	40
ROUNDNESS2	2	7	0	0	0	1	0	1	0	0	11
SOL	2	3	3	0	0	0	0	0	0	0	8
DAE Total	27	28	20	28	18	10	13	24	28	31	227

Table 4.3: Number of Candidate QTLs - Model without Covariables

Trait/DAE	0	1	2	3	4	5	6	7	8	9	Trait Total
COMPACTNESS	12	3	3	5	2	5	1	6	6	3	46
ECCENTRICITY	0	0	0	0	0	0	1	0	0	0	1
ISOTROPY	0	4	0	0	3	2	1	1	1	1	13
PC1	0	0	0	0	0	0	1	1	3	6	11
PC2	4	2	3	2	3	1	1	6	4	5	31
PC3	1	1	0	5	1	0	0	0	0	1	9
PC4	0	0	0	0	2	3	0	0	0	0	5
PC6	0	0	0	1	0	0	0	1	4	1	7
PC7	1	1	0	0	0	1	0	0	0	0	3
PC8	0	0	1	0	0	0	3	1	0	0	5
PC9	1	2	3	2	1	1	0	0	0	0	10
PERIMETER.MM	0	1	2	1	3	2	4	3	3	4	23
ROUNDNESS	5	5	6	3	3	2	4	4	6	5	43
ROUNDNESS2	2	3	0	0	0	1	0	0	0	0	6
SOL	1	1	2	0	0	0	0	0	0	0	43
DAE Total	27	23	20	19	18	18	16	23	27	26	217

Table 4.4: Number of Candidate QTLs - Model with ER_475 as covariable

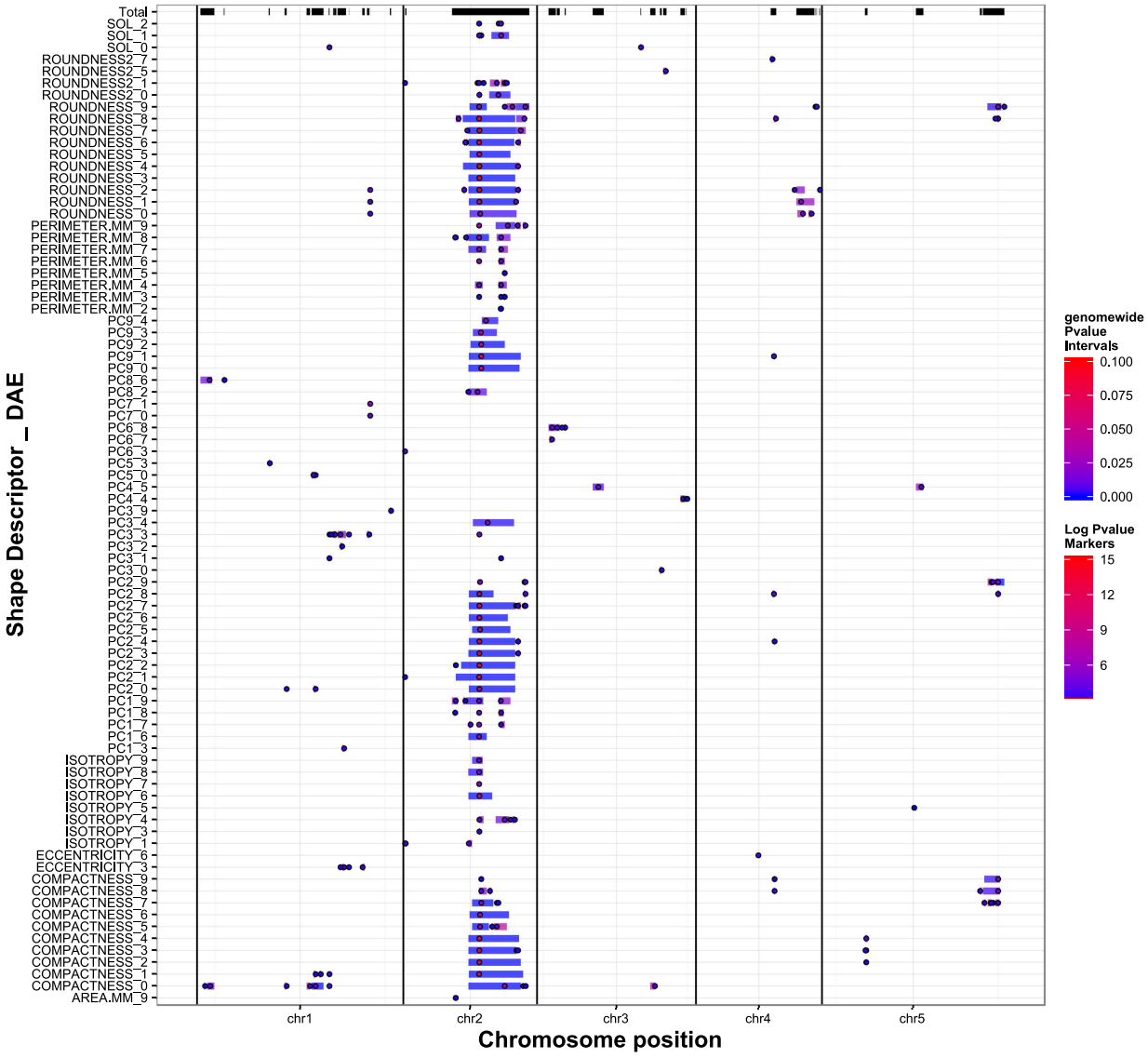


Figure 4.8: QTLs associated to Shape Descriptors - No Covariables

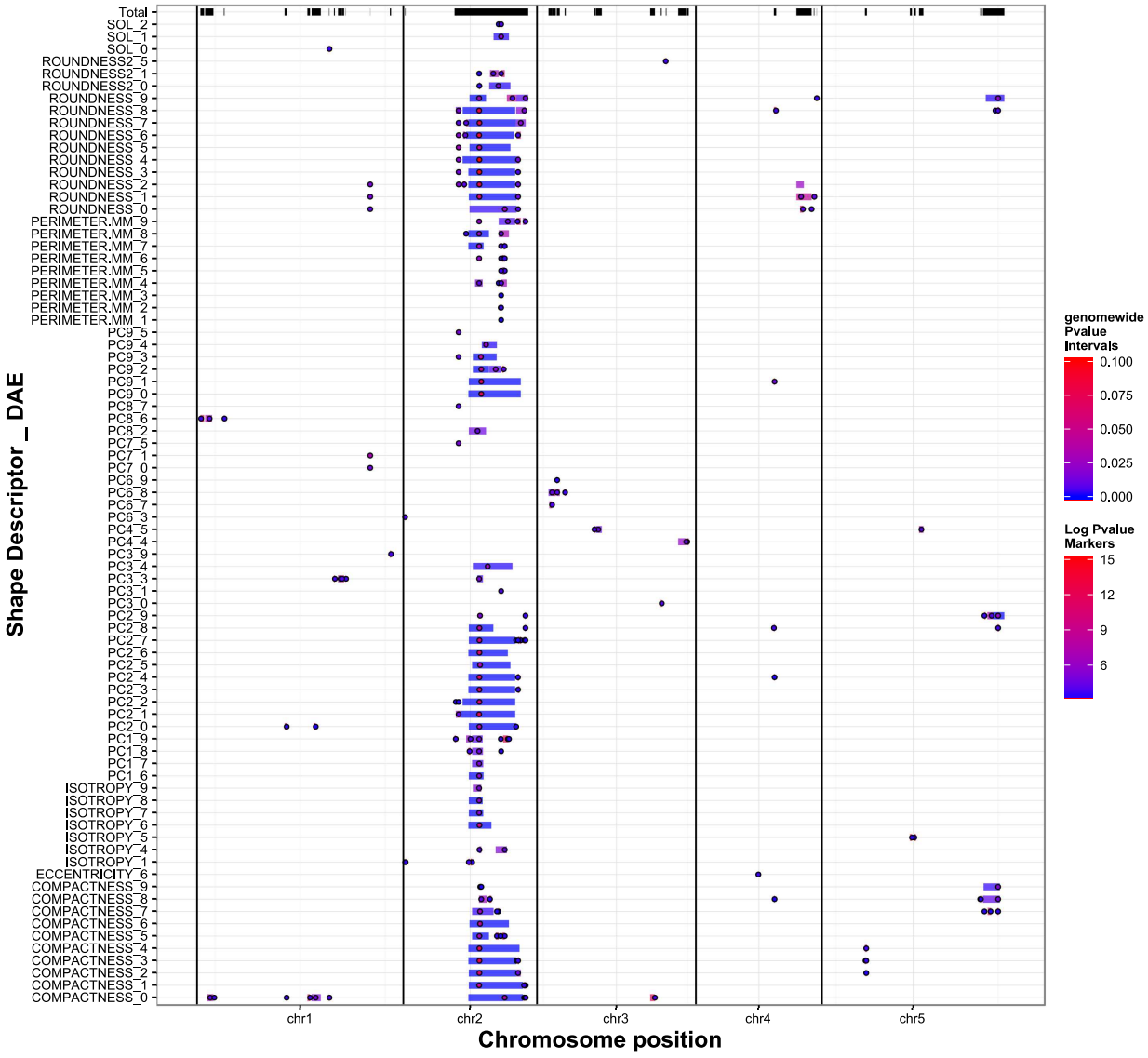


Figure 4.9: Number of candidate QTLs - Using ER_475 as Covariable

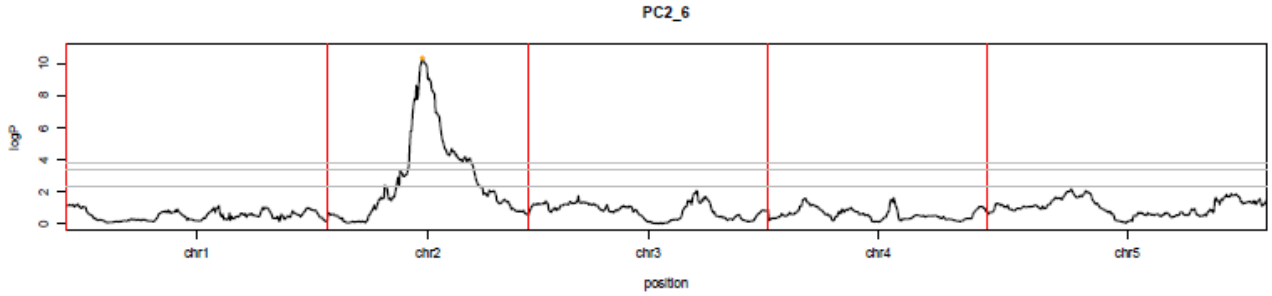


Figure 4.10: QTL profile - PC2 on DAE 6. The marker peak in chromosome 2 corresponds to ER_472

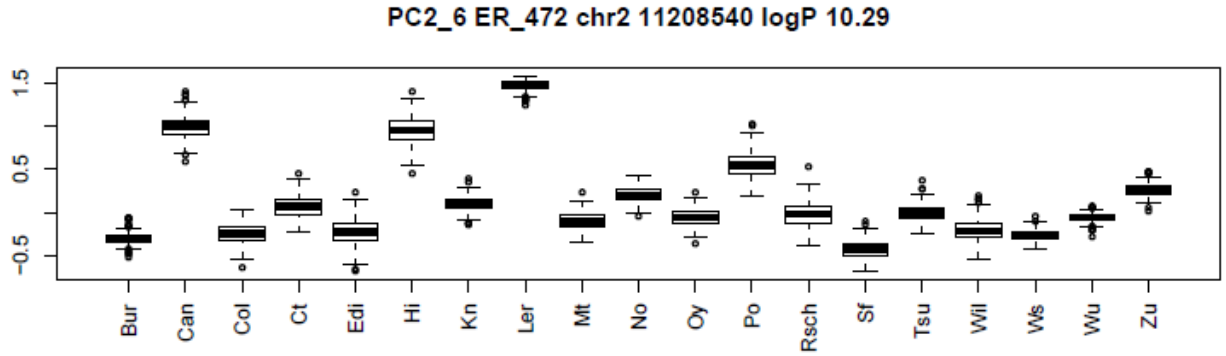


Figure 4.11: Parental of origin effects of marker ER_472 on Principal Component 2 on day 6

Plant.Info	Area.mm	Compactness	Eccentricity	Isotropy	Perimeter.mm	RMS	Roundness	Roundness2	SOL	Founder
MAGIC.16	253.86	0.87	0.15	0.87	88.76	0.46	0.40	0.94	3.87	Ler
MAGIC.164	573.15	0.81	0.19	0.81	148.43	0.44	0.33	0.90	14.61	Can
MAGIC.317	277.91	0.63	0.13	0.73	145.41	0.36	0.17	0.91	8.48	Hi
MAGIC.119	542.29	0.70	0.19	0.53	179.34	0.27	0.21	0.90	25.92	Col
MAGIC.507	164.05	0.77	0.21	0.83	82.45	0.60	0.30	0.92	13.84	Bur
MAGIC.522	180.88	0.58	0.27	0.61	164.89	0.40	0.08	0.83	68.63	Sf

Table 4.5: Shape Descriptor values for selected plants due to the founder of origin of ER_472 alleles

Plant.Info	FOUNDER	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
MAGIC.16	Ler	-1.32	4.79	0.78	0.70	-0.18	0.04	0.50	0.11	0.45
MAGIC.164	Can	-0.66	3.18	0.94	0.32	0.54	0.02	0.34	-0.05	0.23
MAGIC.317	Hi	0.31	1.12	-1.47	0.36	-0.11	0.19	-0.07	0.06	-0.05
MAGIC.119	Col	0.80	0.98	0.15	1.65	0.49	0.46	-0.12	0.26	-0.13
MAGIC.507	Bur	-1.46	2.81	0.40	0.12	-0.67	0.04	0.21	0.13	0.13
MAGIC.522	Sf	0.06	-1.79	-0.15	-0.08	0.47	1.82	-0.24	-0.38	-0.10

Table 4.6: Principal component values for selected plants due to the founder of origin of ER_472 alleles

Figure 4.12: Selection of 6 plants based on their parental of origin for the marker ER_472, written in parenthesis. Tables 4.5 and 4.5 contain the phenotypic values for these plants. The top row contain the RILs with parental of origin that showed higher values for PC2 at DAE 6. Bottom row are RILs with parental of origin that shower low values at the same trait and DAE. The example of RIL 317 is an example whose value is in the middle of high and low value, as well as the shape is intermediated.

4.3.3 Candidate Genes

The extraction of genes contained within the 41 potential QTLs could guide more detailed research into the potential genes related with rosette morphology. As said above, these 41 genome segments are defined by a peak marker and several other markers surrounding it.

The web portal TAIR contain a set of resources to investigate Arabidopsis thaliana genome. The markers used to genotype the MAGIC population are contained in their database system. Thus, looking for markers or chromosome regions in the resources available at TAIR allow to get and insight into potential candidate genes for further studies.

Genes under peak SNPs

Each peak marker, e.g. ER_472, were searched in TAIR and the result were visually inspected for genes that could be related with shape phenotype, growth control or similar. Some possibly related genes are described hereafter. A chromosome map of all genes under peak SNP and the gene description obtained from TAIR are represented in figure 4.13 and table B.3. The segments are identified as Interval X, the number of the interval are in order in the chromosome, but more specific regions are provided at the table B.3. The descriptions are summarized from TAIR "gene search" results

- Interval 13, at chromosome 1, show high significance (logP >6) at the marker FKF1_606 for roundness and PC7 at DAE 0,1 and 2. This marker is located on top of the gene FKF1 (AT1G68050), which is related with transition to flowering, but no phenotype has been associated with variation in this locus.

- Interval 15, at chromosome 2, showed a moderated significativity (logP 3.94) at marker RGA_1023. This SNP is located on top of the gen RGA1 (AT2G01570). This gene is involved in the repression of vegetative growth, induced by Gibberelic Acid, and floral initiation. Some experiments has associated mutants in this gene with dwarf or semi-dwarf phenotypes

- The interval 17, at chromosome 2, maximum logP ∼6.6, was significant for 15 traits and days. Mainly, roundness and PC2, at a range of days from 1 to 7. Three of the 'peak markers' were on the gene PHYB (AT2G18790). This gene codes for a red/far-red photorecep-

tor involved in several process in the cell. It is a well known protein related with petiole elongation and small leaf area. in response to light and shade avoidance.This interval also contains the marker FDP_733 situated on top of the gene FD paralog (AT2G17770) not related with any morphological phenotype, but it is a transcription factor involved in the transition to flowering.

- Interval 18, located at chromosome 2, is the longest interval mapped. Its length is 9.9E6 pair base, around 50% of chromosome 2 length. This interval contains ER_472, which is the marker on top of the gene ERECTA (AT2G26330). Erecta has been related with the absence of trichomes, short petioles, a compact rosette, narrow leaves. In this analysis, variation in the SNP ER_475 has been used as covariate, yet, the logP-value for this marker remain between the highest, e.g ∼12 for roundness at DAE 7. This marker is related to isotropy, roundness and PC2.

  In the same interval, other markers are found significantly related with Shape Descriptors. Roundness at DAE 2 to 8 are associated with two markers, MASC05920 and MASC05927 with highest logP-values of ∼15 and 12 respectively. The same markers showed association with compactness and PC2. MASC05920 is located on top of the gene AT2G26300, coding for an alpha subunit of a protein G. Mutants of this gene has been shown to present reduce cell divisions, round leaves and short hypocotyles. MASC05927 is located on top of the gene AT2G26240, but no phenotype related with plant architecture has been reported yet. The marker MN2_11300378, on top of the gen HO2 (AT2G26550), may have some phenotypes related with the rosette abnormalities in red or far-red lightAll other markers in this interval do not have any curated phenotype associated with plant morphology or architecture.

- Interval 30 consist in a single marker, PHYD_2290, on top of PHYD gene at chromosome 4. The gene PHYD encodes for a phytochrome similar to PHYB, but it has not phenotype associated with rosette structure. Markers' logP-value is relatively small, around 3.5.

- Interval 31 at chromosome 4 contains 3 markers on top of the same gene, AT4G21820, with logP-values around 4. This gene encodes for a calmodulin binding protein expressed in the chloroplast. However, no phenotype related with it has been found.

- Finally, three markers are on top of genes related with porphyrine metabolism, possible related with senescence. They were the markers FKF1_606 (interval 13 at chromosome 1), MNSNP4_15765120 (interval 32 at chromosome 4) and MN2_11300378 (interval 18 at chromosome 2). These markers are on top of the genes AT1G680580, AT4G32690 and AT2626550 respectively. They could be related with senescence as well as with photomorphogenesis. The first two were associated to roundness and PC7 on DAE 0 to 2. The latter one was associated with compactness (DAE 6 and 7) and PC2 (DAE 5 and 9)

Genes On Intervals

A broader search of possible candidate genes consisted on search for genes overlapping the 41 intervals, on the annotated genome of Arabidopsis thaliana release 9 (TAIR9). The genes found in this search were compared to a set of shape related genes found by Wilson-Sánchez et al. (2014) and released in the Phenoleaf database. The 41 intervals where split in two sets, on one side the big interval at chromosome 2, and on the other side all the other intervals.

The 40 intervals overlapped with 5667 genes in the TAIR 9 annotated genome. The big interval by itself alone overlaps with 2910 other genes in the same genome.

The intersection of these genes with PhenoLeaf database is represented in figure 4.14 and tables 4.7 and 4.8. The set genes on the 40 intervals overlaps with 90 genes in the PhenoLeaf dataset (table 4.7), while the genes in the big interval overlaps with 49 genes in that dataset (table 4.7). This make a total of 139 candidate genes whose variation may be affecting the shape of our MAGIC RILs population.

Gene ID	Description
AT1G02480	pre-tRNA. tRNA-Phe (anticodon: GAA)
AT1G02670	Protein of unknown function
AT1G03020	Thioredoxin superfamily protein
AT1G03070	Bax inhibitor-1 family protein
AT1G04730	CHROMOSOME TRANSMISSION FIDELITY 18 (CTF18)
AT1G06320	Protein of unknown function
AT1G06340	Plant Tudor-like protein

Table 4.7: Set on genes within 40 QTL intervals included in PhenoLeaf

Gene ID	Description
AT1G07650	Leucine-rich repeat transmembrane protein kinase
AT1G07660	Histone superfamily protein
AT1G11790	AROGENATE DEHYDRATASE 1 (ADT1)
AT1G35612	Pseudogene of Ulp1 protease family protein
AT1G43690	Ubiquitin interaction motif-containing protein
AT1G47630	CYTOCHROME P450, FAMILY 96, SUBFAMILY A, POLYPEPTIDE 7 (CYP96A7)
AT1G47813	Protein of unknown function
AT1G47890	RECEPTOR LIKE PROTEIN 7 (RLP7)
AT1G48750	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
AT1G48950	C3HC zinc finger-like
AT1G49210	RING/U-box superfamily protein
AT1G52240	RHO GUANYL-NUCLEOTIDE EXCHANGE FACTOR 11 (ROPGEF11)
AT2G18790	PHYTOCHROME B (PHYB)
AT3G06270	Protein phosphatase 2C family protein
AT3G07610	INCREASE IN BONSAI METHYLATION 1 (IBM1)
AT3G08040	FERRIC REDUCTASE DEFECTIVE 3 (FRD3)
AT3G08920	Rhodanese/Cell cycle control phosphatase superfamily protein
AT3G09720	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT3G23660	Sec23/Sec24 protein transport family protein
AT3G25470	Bacterial hemolysin-related
AT3G25520	RIBOSOMAL PROTEIN L5 (ATL5)
AT3G25585	AMINOALCOHOLPHOSPHOTRANSFERASE (AAPT2)
AT3G25740	METHIONINE AMINOPEPTIDASE 1C (MAP1B)
AT3G46230	HEAT SHOCK PROTEIN 17.4 (HSP17.4)
AT3G46610	Pentatricopeptide repeat (PPR-like) superfamily protein
AT3G46790	CHLORORESPIRATORY REDUCTION 2 (CRR2)
AT3G49040	F-box/RNI-like superfamily protein
AT3G49190	O-acyltransferase (WSD1-like) family protein
AT3G49460	Protein of unknown function
AT3G56170	CA-2+ DEPENDENT NUCLEASE (CAN)
AT3G57390	AGAMOUS-LIKE 18 (AGL18)

Table 4.7: Set on genes within 40 QTL intervals included in PhenoLeaf

Gene ID	Description
AT3G57810	Cysteine proteinases superfamily protein
AT3G57940	Putative ATPase
AT3G58360	TRAF-like family protein
AT3G58960	F-box/RNI-like/FBD-like domains-containing protein
AT3G59460	Similar to F-box family protein
AT3G60240	EUKARYOTIC TRANSLATION INITIATION FACTOR 4G (EIF4G)
AT4G30410	Sequence-specific DNA binding transcription factor
AT4G30540	Protein of unknown function
AT4G30580	LYSOPHOSPHATIDIC ACID ACYLTRANSFERASE 1 (LPAT2)
AT4G31110	Protein of unknown function
AT4G31390	Protein kinase superfamily protein
AT4G31490	Contains domain Coatomer, beta subunit
AT4G31530	NAD(P)-binding Rossmann-fold superfamily protein
AT4G31990	ASPARTATE AMINOTRANSFERASE 5 (ASP5)
AT4G32040	KNOTTED1-LIKE HOMEOBOX GENE 5 (KNAT5)
AT4G32105	Beta-1,3-N-Acetylglucosaminyltransferase family protein
AT4G32200	ASYNAPTIC 2 (ASY2)
AT4G32670	RING/FYVE/PHD zinc finger superfamily protein
AT4G32780	Phosphoinositide binding
AT4G32810	CAROTENOID CLEAVAGE DIOXYGENASE 8 (CCD8)
AT4G32930	Protein of unknown function
AT4G33520	P-TYPE ATP-ASE 1 (PAA1)
AT4G34000	ABSCISIC ACID RESPONSIVE ELEMENTS-BINDING FACTOR 3 (ABF3)
AT4G34730	Ribosome-binding factor A family protein
AT4G35650	ISOCITRATE DEHYDROGENASE III (IDH-III)
AT4G36870	BEL1-LIKE HOMEODOMAIN 2 (BLH2)
AT5G36880	ACETYL-COA SYNTHETASE (ACS)
AT5G58170	SHV3-LIKE 5 (SVL5)
AT5G58670	PHOSPHOLIPASE C1 (PLC1)
AT5G58970	UNCOUPLING PROTEIN 2 (UCP2)
AT5G60410	SIZ1

Table 4.7: Set on genes within 40 QTL intervals included in PhenoLeaf

Gene ID	Description
AT5G60790	ATP-BINDING CASSETTE F1 (ABCF1)
AT5G61050	Histone deacetylase-related / HD-related
AT5G61170	Ribosomal protein S19e family protein
AT5G61240	Leucine-rich repeat (LRR) family protein
AT5G61950	Ubiquitin carboxyl-terminal hydrolase-related protein
AT5G61960	MEI2-LIKE PROTEIN 1 (ML1)
AT5G62190	DEAD/DEAH box RNA helicase PRH75
AT5G62290	Nucleotide-sensitive chloride conductance regulator (ICln) family protein
AT5G62660	F-box and associated interaction domains-containing protein
AT5G62800	Protein with RING/U-box and TRAF-like domains
AT5G63820	Protein of unknown function
AT5G64080	XYLOGEN PROTEIN 1 (XYP1)
AT5G64710	Putative endonuclease or glycosyl hydrolase
AT5G64790	Protein of unknown function
AT5G65050	AGAMOUS-LIKE 31 (AGL31)
AT5G65420	CYCLIN D4.1 (CYCD4.1)
AT5G65640	BETA HLH PROTEIN 93 (bHLH093)
AT5G66020	SUPPRESSOR OF ACTIN 1B (ATSAC1B)
AT5G66330	Protein of unknown function
AT5G66490	Protein of unknown function
AT5G67270	END BINDING PROTEIN 1C (EB1C)

Table 4.7: Set on genes within 40 QTL intervals included in PhenoLeaf

Gene ID	Description
AT2G19790	SNARE-like superfamily protein
AT2G20270	Thioredoxin superfamily protein
AT2G20560	DNAJ heat shock family protein
AT2G20580	26S PROTEASOME REGULATORY SUBUNIT S2 1A (RPN1A)
AT2G21660	COLD, CIRCADIAN RHYTHM, AND RNA BINDING 2 (CCR2)

Table 4.8: Set on genes within “large” QTL interval at chromosome 2 included in PhenoLeaf

Gene ID	Description
AT2G22090	UBP1-ASSOCIATED PROTEIN 1A (UBA1A)
AT2G22420	Peroxidase superfamily protein
AT2G22460	Protein of unknown function
AT2G22470	ARABINOGALACTAN PROTEIN 2 (AGP2)
AT2G22540	SHORT VEGETATIVE PHASE (SVP)
AT2G22680	Zinc finger (C3HC4-type RING finger) family protein
AT2G23090	Uncharacterised protein family SERF
AT2G23220	CYTOCHROME P450, FAMILY 81, SUBFAMILY D, POLYPEPTIDE 6 (CYP81D6)
AT2G23380	CURLY LEAF (CLF)
AT2G23840	HNH endonuclease
AT2G24090	Ribosomal protein L35
AT2G25870	Haloacid dehalogenase-like hydrolase family protein
AT2G27530	PIGGYBACK1 (PGY1)
AT2G28450	Zinc finger (CCCH-type) family protein
AT2G28725	Protein of unknown function
AT2G29300	NAD(P)-binding Rossmann-fold superfamily protein
AT2G29670	Tetratricopeptide repeat (TPR)-like superfamily protein
AT2G30170	Protein phosphatase 2C family protein
AT2G30810	Gibberellin-regulated family protein
AT2G31650	HOMOLOGUE OF TRITHORAX (ATX1)
AT2G31725	Eukaryotic protein of unknown function
AT2G31870	SANSKRIT FOR BRIGHT (TEJ)
AT2G32540	CELLULOSE SYNTHASE-LIKE B4 (CSLB04)
AT2G32760	Protein of unknown function
AT2G33030	RECEPTOR LIKE PROTEIN 25 (RLP25)
AT2G33050	RECEPTOR LIKE PROTEIN 26 (RLP26)
AT2G33250	Protein of unknown function
AT2G33420	Protein of unknown function, contains domain Munc13 homology 1
AT2G33450	Ribosomal L28 family
AT2G34260	HUMAN WDR55 (WD40 REPEAT) HOMOLOG (WDR55)
AT2G35040	AICARFT/IMPCHase bienzyme family protein

Table 4.8: Set on genes within “large” QTL interval at chromosome 2 included in PhenoLeaf

Gene ID	Description
AT2G35720	ORIENTATION UNDER VERY LOW FLUENCES OF LIGHT 1 (OWL1)
AT2G36480	ENTH/VHS family protein
AT2G36620	RIBOSOMAL PROTEIN L24 (RPL24A)
AT2G37290	Ypt/Rab-GAP domain of gyp1p superfamily protein
AT2G37650	GRAS family transcription factor
AT2G38330	MATE efflux family protein
AT2G40650	PRP38 family protein
AT2G40750	WRKY DNA-BINDING PROTEIN 54 (WRKY54)
AT2G40940	ETHYLENE RESPONSE SENSOR 1 (ERS1)
AT2G41140	CDPK-RELATED KINASE 1 (CRK1)
AT2G41680	NADPH-DEPENDENT THIOREDOXIN REDUCTASE C (NTRC)
AT2G42720	FBD, F-box, Skp2-like and Leucine Rich Repeat domains containing protein
AT2G44100	GUANOSINE NUCLEOTIDE DIPHOSPHATE DISSOCIATION INHIBITOR 1 (GDI1)

Table 4.8: Set on genes within "large" QTL interval at chromosome 2 included in PhenoLeaf

To refine the search for candidate genes, the set of genes underlying peak markers were matched with those on PhenoLeaf database. Only the genes AT2G18790, AT2G22680 and AT2G42720 where common in both sets.

AT2G18790 is under markers PHYB_2850, PHYB_4171 PHYB_5215 and correspond to the phytochrome B gene. AT2G22680 is under the marker MN2_9653239 and correspond to the gene WAV3 Homolog. This gene encodes for a protein with activity ubiquitin-protein ligase and zinc ion binding. This protein does not have a shape-related phenotype in the Phenoleaf database, but it does have a phenotype related with the netted leaf color pattern. The gene AT2G42720 is under the marker MN2_17792406. It encodes for a FDB, F-box, Spd-2 like protein. According to Phenoleaf database it affects to rosette compactness and the roundness of individual leaves.

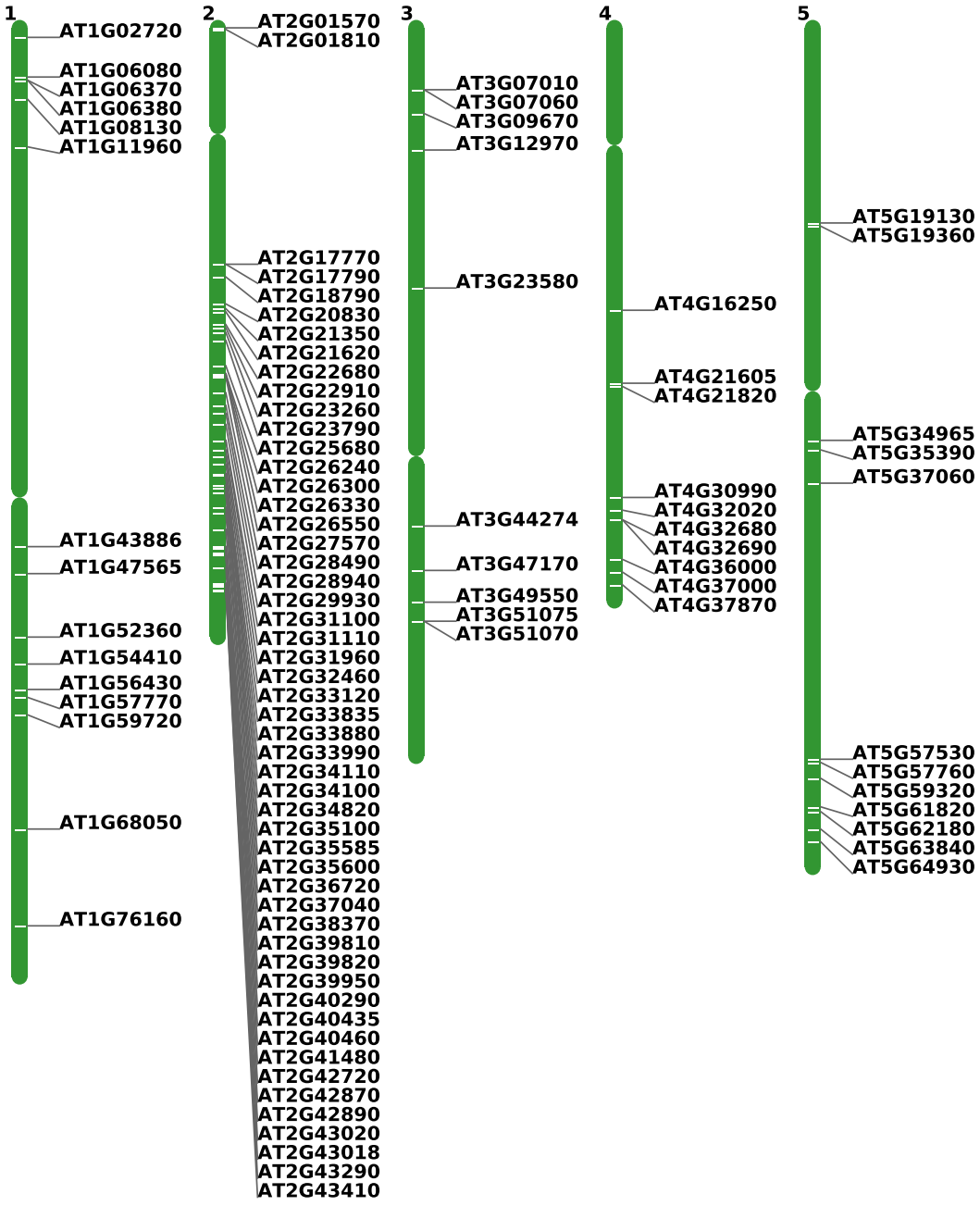


Figure 4.13: QTLs associated to Shape Descriptors - Using ER_475 as Covariable. Chromosome Map

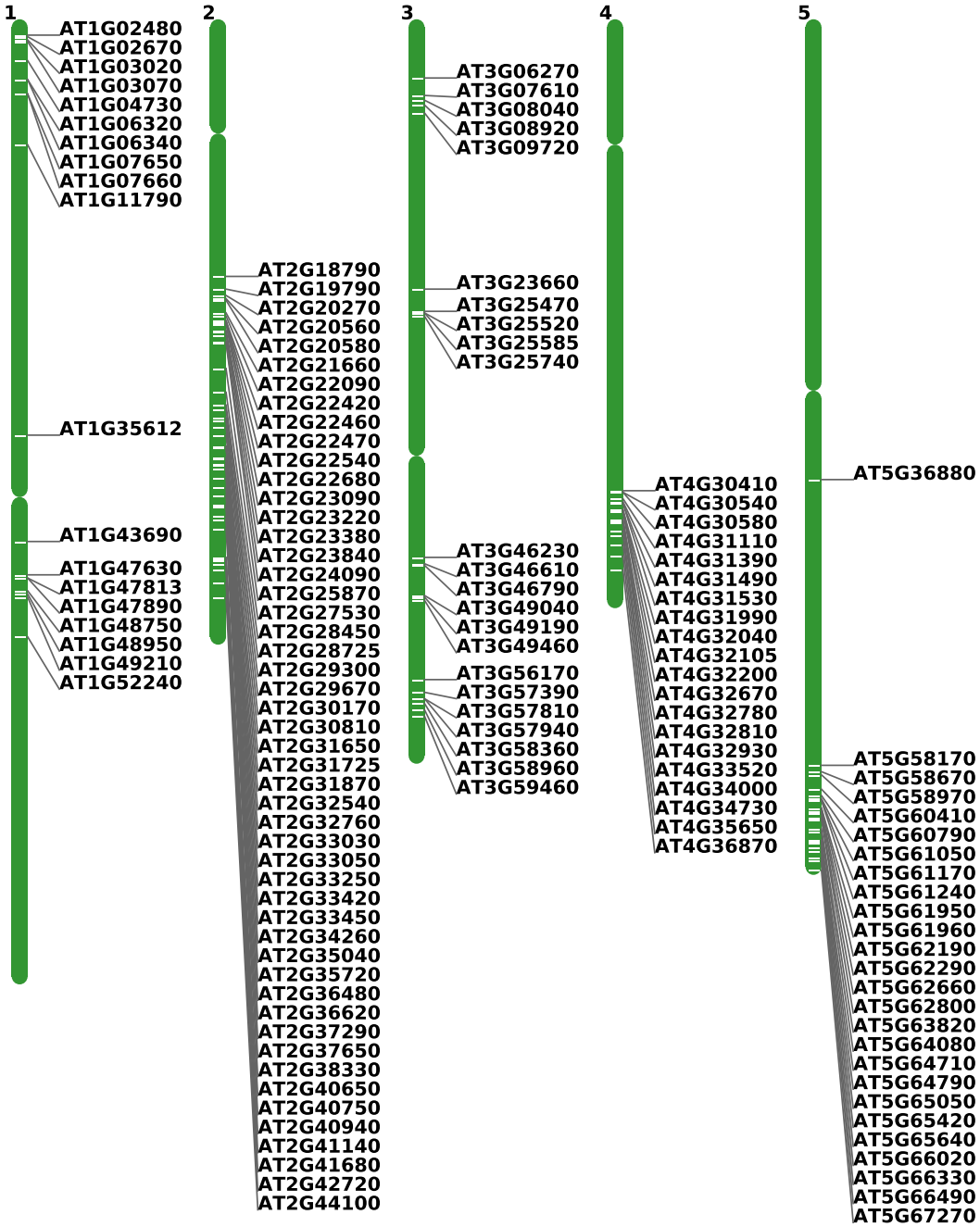


Figure 4.14: QTLs associated to Shape Descriptors - Using ER_475 as Covariable. Intersection of qShape intervals and PhenoLeaf dataset. Chromosome Map

Gene Enrichment

Gene enrichment test was performed to check whether genes found within intervals contains more genes from PhenoLeaf database than expected by chance. Phenoleaf is a database of mutant leaf phenotypes in the Arabidopsis Salk Unimutant collection, and putative responsible genes resulting from screening tests. The gene enrichment was done using an hypergeometrical test.

All intervals together sum up to 8681 genes, 139 from Phenoleaf and 8542 not contained in Phenoleaf. Based on 33239 genes in the annotated genome from TAIR9, it is expected 133 genes from Phenoleaf. This mean a fold enrichment of 1.04 with a p-value in the hypergeometrical test of 0.27.

5667 genes were within all intervals except the big one on chromosome 2, and 90 genes were contained in Phenoleaf (87.12 genes expected) representing a fold enrichment of 1.03 (p-val 0.34).Within the big interval, there were 2910 genes, being 49 genes in the Phenoleaf (44.74 expected) representing an enrichment of 1.09 (p-value 0.22).

Considering only the genes under peak markers, 217 genes, the 3 genes found in the PhenoLeaf dataset represent a fold-enrichment of 0.89 (3.33 genes expected, p-value 0.42).

Thus all the enrichment test indicates that genes on the PhenoLeaf dataset were present as expected by chance.

Another Gene Enrichment test has been performed using the GO ontology database AMIGO (`http://amigo.geneontology.org/amigo`). Ontology search for molecular function, biological process and "PANTHER Pathways" were performed using Bonferroni correction for multiple test.

For the set of genes under peak markers, in the category molecular functions only the ontology 'photoreceptors' showed an enrichment of 63.71 (3 genes out of 0.5 expected, pval = 2.26E-2). No significant enrichment was found under any other category.

For comparison purpose, the same analysis was performed on the original PhenoLeaf database set of genes. For molecular function, the only category significantly enriched was "binding" with an enrichment value 1.28 (203 genes out of 159.18 expected, pval = 2.71E-2). For biological process, the category 'cellular response to stimulus' got an enrichment factor of 1.72 (63 genes out of 36.64 expected, pval =4.62E-2). The subcategory 'cellular processes) show an 1.27

enrichment factor (206 genes out of 161.62 expected, pval = 3.58E-2). The category 'single organism cellular process' was 1.51 enriched (131 genes out of 86.69 expected, pval = 9.14E-4). No category from 'Panther Pathways' were significantly enriched.

Similar analysis was performed for the genes within the 41 intervals. The first analysis pools together the 40 intervals, excluding the big interval at chromosome 2. For molecular function, no significant positive enrichment was found in this set, but a negative enrichment in the category 'Oxygen Binding' of 0.34 (12 genes out of 35.26 expected, pval = 7.31e-3) appear. No significant enrichment was found for biological processes and "PANTHER pathways".

For the big interval at chromosome 2, enrichment was again found only for molecular function ontologies. Positively enrichment was found the category 'Chitin binding' with a factor of 6.85 (8 genes out of 1.17 expected, pval = 4.21E-2) and a negative enrichment of 0.2 in the category 'ADP binding' (1 gene out 14.11 expected, pval= 1.51E-2). However, the latter values may be due to the single location of this interval in the chromosome 2, so the assumption of independence in the sample drawn is not fulfilled.

4.4 Discussion

The correlation analysis of shape descriptor on the MAGIC population revealed the correlation of shape with size. Area and perimeter can be considered as pure size measurements, containing almost no information on shape by themselves. Thus, the correlation between other shape descriptor with area and perimeter reveals the aforementioned relationship between size and shape. A particular example is the correlation of slender of leaves with perimeter. It indicates that most of the information contained in this descriptor is brought by the perimeter. This happen even when actual perimeter is not included in the formula of the SOL descriptor. The analysis of correlations between descriptors induce to think shape descriptors as different metrics of same concept of shape. Compactness is a strong, repeatable phenotype by itself, it accounts for the filling of the space by the rosette. Shape Descriptors accounting for similarities between rosettes and circles (roundness, roundness2 and isotropy), ellipses (eccentricity) can be thought, due to their correlation, and the conceptualization underlying them, as redundant measurements of rosette spatial distribution. Slender of leaves, although designed to measure the distribution of petioles and blades length, it is redundant with roundness and isotropy.

Thus, I consider that these shape descriptors are similar enough to be interpreted as collinear measurements of the same characteristic. This justify to melt the results of candidate QTLs into "qShape" or QTL for Shape, independently of the specific shape descriptor that helped to discover it.

Principal Component Analysis (PCA) may help remove the influence of size from shape and get pure shape measures (Humphries et al., 1981; Sundberg, 1989; Somers, 1989). PCA split shape descriptors into 9 orthogonal variables, i.e not correlating. The first Principal Component (PC) is composed mainly of area, perimeter and SOL, being the component with most of the size related parameters.The second Principal Component does not contain neither area nor perimeter, keeping most of rosette shape. The other PCs were included in the QTL mapping, but they were not a clear measurement of shape.

According to heritabilities, only area, perimeter, compactness and roundness are heritable through time, although decreasing as experiment progress. This convey the difficulty of discriminate "qShape" as spurious associations, i.e False Positives, or actual genetic determinants. Similar pattern of variability in broad heritability along time has been reported by Flood et al. (2016) . According to these authors, such displacement in heritability may be due to the frequency of measurement, so this effect should not be detected at frequencies of single points per day (Flood et al., 2016). Their study is based on photosynthetic parameters, but it may be still true for developmental trajectories of rosette shape.

Principal Components 1, 2 and 3 had heritabilities high enough to accept that they contain enough phenotypic variation explained by genetic variation. The heritability values show a descending tendency through time, with their highest values around day 1 and 2.

The statistical association between markers and traits result in 41 candidate Quantitative Trait Loci along the genome. This suggest a complex regulation of the global rosette shape. To ascertain the real genetic factors hidden under those loci is not possible given only significantly associate markers and their associate regions. Nevertheless, these loci are candidate regions to be researched by fine mapping in near-isogenic lines or other experimental populations.

In this analysis, 217 markers in 41 intervals were found as associated to rosette shape. As a comparison, Chitwood et al. (2013b) studied tomato leaf shape and found 1000 QTLs. Chitwood et al. (2013b) claimed that such numbers suggested additive polygenic effect all

over the genome with a limited role for epistasis. In addition, the authors suggested that tomato leaves where influenced by QTLs that regulates mostly cellular phenotypes according to environmental abiotic conditions (Chitwood et al., 2013b).

The presence of genes within our QTLs that are related with either transition to flowering, senescence or photomorphogenesis, suggest that the revealed potential QTLs for shape may not be false discoveries. During the experiments, differential flowering time and senescence was observed, and it is feasible that the shape descriptors utilized vary according to this phenomena.

The most relevant group of genes, to my opinion, are those related with photomorphologenesis and adaptation to environmental clues. The phytochromes B and D, are well known responsible of leaf growth changes according to shade-avoidance patterns (Tsukaya, 2004; Kozuka et al., 2005). The gene *phyB* is responsible of petiole elongation in the search of light (Tsukaya et al., 2002; Tsukaya, 2004). The gene *erecta*, is known to be related with Arabidopsis morphology (Passardi et al., 2007), and it has been proposed that either *erecta* or a closely linked gene is a regulator of phenotypic canalisation due to microenvironmental variation (Hall et al., 2007b; van Zanten et al., 2009a).

The enrichment tests showed that shape related genes are not straightforwardly associated to any molecular function, pathway or process, as expected from additive traits such as size. However, our test showed that photoreceptors were found with higher chance than expected, so that it is conceivable to think that environmental interaction, GxE, has played a role in our gene discovery experiment.

The genes within candidate shape QTLs contains some transcription factors and regulatory sensors to environmental cues (table B.3). Few of them were directly related with leaf shape, as suggested by our enrichment analysis against PhenoLeaf database (Wilson-Sánchez et al., 2014). This coincides with recent literature describing how Arabidopsis rosette developmental trajectories are strongly affected by local environmental conditions. Bar and Ori (2014) review leaf development and morphogenesis in the context of environmental effects on the final shape. Mature leaves shape is influenced by light through phytochrome-induced responses, to resource limitation, at least in Arabidopsis, through auxine and gibbereline hormonal response (Bar and Ori, 2014). Salinity seems to act through abcisic acid and ethylene response by the protein DELLA . Many of candidate genes underlying the peak marker in our candidate shape

QTLs belong to these groups. To show some examples contained in table B.3, AT2G01570 is a member of DELLA family, AT2G42870 is PHYTOCHROME RAPIDLY REGULATED1, AT2G33880 is a WUS related protein necessary in meristem growth, AT2G37040 encodes for PAL1 (phenylalanine ammonia lyase) related with photomorphogenesis through PhyA, and with the formation of anthocyanins.

If those genes were causal factors whose variation explain rosette variation, these findings point at regulatory elements connecting rosette architecture to environmental variation.

According to our results, it is reasonable to think that Global Shape Descriptors captures the populational variability and the reaction norm of Arabidopsis rosette architecture. To certain point, instead of variation in shape, we have may found accessions that are more variable than others in its general structure and growth according to the environment they live. Thus, we have, unintentionally, looked for the robustness of their architecture or the phenotypic plasticity and their ontogenic adaptation to the local conditions.

Chapter 5

Discussion

In this thesis, I have put together three technical approaches to dissect *Arabidopsis thaliana* shoot development. First, measurements of overall rosette shape has been proposed and used. Second, High-throughput phenotyping, i.e. measurement, through imaging of rosettes. Third, phenotype to genotype mapping tools have been applied to unveil the genetic architecture underlying rosette structure.

It was initially observed that natural ecotypes of Arabidopsis have different developmental trajectories of growth during their juvenile stage (Camargo et al., 2014). "Shape descriptors", accounting for whole rosette shapes, have been adopted to quantify the diversity and variation of such rosettes, within and between varieties. Variation in shape descriptors, interpreted as morphological traits, have been mapped to genomic positions, meaning that genetic loci accounting for variation in rosette shape has been approximately located. The mapping resolution was not enough in any experiment to identify a causal gene, yet some hypothesis have been generated.

Due to progressive leaf production across time, individual leaf shape and structure, albeit being an interesting problem itself, is of little help to study rosette development due to Arabidopsis heteroblasty (Poethig (2013), Robbelen (1957, cited in Tsukaya (2013)),Tsukaya (2002)). The study of a specific leaf at specific moment, e.g the 6th leaf on the 20th day, does not provide a complete picture of rosette appearance across time. Geometrical morphometrics of whole individuals could be an appropriate approach to study overall rosette shape, with the inconvenience that "landmarks" are difficult to define in a developing rosette, and it is a manual process with many difficulties to be adapted to automatic marking for high-throughput

purposes. The option I found as more feasible was the use of Global Descriptors (Zhang and Lu, 2004). Computer vision has a rich history in the developing such kind of descriptors that can be automatically calculated from a segmented image. They allow quantification of shape of any object in general and Arabidopsis rossetes in particular. However, a major drawback of shape descriptors is that they remain global, without any direct correlation with particular aspects of leaves such as petiole length, blade length, phyllotaxis, etc.

Shape Descriptors can be only calculated from images that have been processed so that rosettes are the only element in the image and the background is removed (Shapiro and Stockman, 2001). This processing is named in the specific literature segmentation (Haralick and Shapiro, 1991). In general, segmentation is a family of problems with difficult resolution, meaning that good quality segmentation is infeasible unless pictures are taken in a very controlled environment, e.g. industrial quality control (Shapiro and Stockman, 2001). In general, shape descriptors and similar techniques are used only as approximations either to evaluate the result of a segmentation or to evaluate the probability whether the object in the image is from the same category that other images in a database, i.e. Image Retrieval (examples at Utku, 2000; Gopal et al., 2012; Harish et al., 2013). However, certain shape descriptors has been used to quantify the properties and structure of compact objects, such as powders, materials, seeds and fruits (Brosnan and Sun, 2004) and debate about their efficacy remains (Pirard and Dislaire, 2005, 2011). However, few examples of shape descriptors for quantify non-compact objects, e.g star-fish like objects or curved objects, have been found in the literature. To my knowledge, only Area and Compactness are of wide use in Arabidopsis phenotyping (see table 1.1). A deeper discussion will be found below.

The recent development of automatic phenotyping devices, such as Scanalyzer (Lemnatec, Gmbh) and PlantScreen (PSI, Czech Republic), has opened the opportunity to study plant development from a dynamic perspective in a high-throughput manner (Li and Sillanpää, 2015). Most of this advance is due to the use of a diverse range of imaging devices, for example fluorescence and Infra-red cameras (Rahaman et al., 2015). With simple visible wavelength cameras, i.e classical colour camera, the shape and colour of plants can be taken to analyse the structure and processes (Sozzani et al., 2014; Dhondt et al., 2014; Humplík et al., 2015). The improvement carried by imaging is that a device can take pictures, and save them in database

for latter analysis, of potentially hundreds of plants, and also can take the pictures a number of times per day (Furbank and Tester, 2011). This supplies researchers with huge amount of information that allows to analyse the measured traits dynamically, time-resolved, in as many points as required.

These automatic platforms, and the analysis now associated with them, facilitate the use of populations with large number of individuals (Fahlgren et al., 2015b). This is ideal for mapping populations whose resolution depends largely on the genetic diversity of the population and the number of informative crossover cumulate in them (Takuno et al., 2012). The more individuals the more likely to shorten the haplotypic regions and more resolution. Similarly, the higher number of replicates the more accurate phenotypic values are measured.

The approach in this thesis has made an extensive use of the power provided by automatic phenotyping platforms to study a similar set of rosette shape descriptors in three mapping populations with different genetic and phenotypic properties.

The experiment in chapter 2 uses a population of natural accessions, also called ecotypes. This population is expected to have large genetic and phenotypic variation that make it suitable for mapping with relative accuracy phenotypes with a complex genetic architecture. The main drawbacks of natural population is that accessions are not independent of each other in the phylogenetic sense. This means that some accessions are genetically and historically closer of each other than they are to some others. The effect of this population structure is having some chromosome regions identical by descent and in linkage disequilibrium, which is a confounding factor in the analysis. A second drawback is that these populations have been subject to evolutionary forces, e.g. adaptation through selection, to cope with their local environments. Although debate still exists about the levels of selection on the gene or the genome, it is generally agreed that populations are selected by their complete phenotype, so that many particular phenotypes, and the genes that control them, are selected simultaneously (Winter, 1997; Haldane, 2008; Pigliucci, 2010). Thus, genes for two different process being selected at the same time would be confounded when doing gene mapping because they coexist in similar linked frequencies (Vilhjálmsson and Nordborg, 2012; Platt et al., 2010; Brachi et al., 2011).

The results from the natural ecotypes population were inconclusive. None of the shape descriptor show significant association to any genetic markers. A procedure to further study

those markers having large LOD (but not significant) for more than 10 traits was chosen to extract the most informative markers, but it is not generally used as a formal approach. The reason for having low significance, in spite that some LOD scores were over 7, generally considered high, is probably the lack of resolution in the images. This first experiment was performed in the Lemnatec device, that was built for tall crops like Miscanthus, maize or wheat. The top-view camera is placed on the roof, around 2 meters on top of the rosettes. The long distance from rosette to camera lens reduces the pixel resolution of rosettes. As an example, typical petioles were just one pixel width. In addition, the camera objective in Lemnatec device endure chromatic aberrations, i.e. any object in the picture presents a rainbow-like gradient of colours in their borders, that affect strongly to the segmentation. In spite of the technical difficulties, at least four potential Quantitative trait loci seems to be associated with the shape descriptors. However, a search of intervals of $\pm 10kb$ around the markers did not yield any gene that were already studied in the literature as related with Arabidopsis growth, neither rosettes or leaves.

The experiment in chapter 3 used a biparental cross between two accessions that showed disparate rosette development in chapter 2. Biparental crosses have no population structure, but they have less genetic variation. The drawback of these kind of population is that only genes that differ in alleles in both parental can show any degree of significance in the association with the trait. If selection has fixed alleles for major regulators in the trait of interest, only minor effect genes can be found, and those need bigger populations. Our population is relatively small and with few genotyped markers, allowing to find only major effect genes.

Despite these limitations, the results indicate that at least variation in 4 QTL for shape were present in this population, and they were not far from those found in chapter 2. The populations in chapter 2 and in chapter 3, were genotyped using different markers and therefore have different maps, it is not possible to establish any formal comparison between regions.

In chapter 4, I used a 19 parent cross, the resource is called MAGIC, that is expected to have levels of genetic variation between natural and biparental crosses. Also little or no population structure is expected.

The results obtained from the MAGIC population were richer than in the previous chapters. Each descriptor had significant association with many markers in any chromosome. I decide

to unify all intervals that overlap to reduce the search. This decision was taken after noticing that the association mapping method developed for MAGIC does not calculate intervals as the classical confidence intervals in genetic mapping, but they are just the extent of significant markers. In other words, the method returns as an interval all the markers that were significantly associated to a phenotype. This is due to the method related to haplotype mapping, so all markers in the same haplotype has a similar significance. Using this joint approach around 40 intervals, potential QTLs were found. A search for genes within these intervals returned too many genes, meaning the search was like "looking for a needle in a haystack". These genes were screened against a database, PhenoLeaf (Wilson-Sánchez et al., 2014), of genes related with leaf shape from an experiment done with a populations of mutants, and also were studied from a gene enrichment approach. Both approaches indicate that the number of genes found were the number expected by chance. This means that I cannot claim any of those particular genes as potential causal genes.

It could be argued whether the MAGIC intervals contain or not the QTLs found in chapter 2 and 3. The big interval in the chromosome 2 is the joint of two clear groups of markers at the middle of the chromosome. This may or not be related with the potential QTL at this chromosome found in the natural ecotypes population. A potential QTL was also found at the end of chromosome 4, being more probable to be in a close position of those found in chapter 2 and 3, since the interval in the chromosome 4 did not extent long neither in the MAGIC nor in the natural population.

The summary from the three experiments is that shape descriptors has not been found clearly associated to any particular loci in the genome. Many potential candidate QTLs have emerge from these experiments, but not with the precision to point to a particular candidate gene.

This is expected, since shape, like size, could be controlled by many genes with complex interactions, being few of them major controllers that are probably acting as integrators of many pathways (Martin et al., 2011). However, the work performed here act as observation for future research. It might be argued that genes in the middle of the chromosome 2, at the end of chromosome 3, at the middle of 4 and at the end of chromosome 5 have been observed as significant, or almost, in the three studies performed here. Assuming any explanation from

this observation would be "*ad hoc*", however, it is know that major regulators of life-trait and leaf shape are in those regions. The gene *FLC* that regulates developmental pathways (Deng et al., 2011) is at the end of chromosome 5. The gene *ERECTA* is in the middle of chromosome 2. In addition phytochromes B and D has been found as potential QTLs. *PHYB* is located in the middle of chromosome 2 and *PHYD* in the middle of chromosome 4. Phytochromes are related with the rate of growth of leaves and the environmental response to light and the shade avoidance syndrome (Martínez-García et al., 2014), a phenotypic plasticity response to competition for light (Schmitt et al., 1999). They participate in the elongation of petioles and also as thermosensors (Jung et al., 2016). For these reasons, phytochromes and *erecta* are genes whose further exploration can be interesting.

Throughout this project, some questions to observations remain unanswered and can be potential future research. Theoretically, if rosette shape, size and development "rythm" are genetically controlled, this could have been the result of developmental canalization in the original accessions environment. This is assumed to happen when development "takes control" of variation due to environmental conditions, instead of passive response, and variation is buffered or even leveraged (see Salazar-Ciudad, 2007, for a discussion). One example is found in Hall et al. (2007a) who studied canalization of rosette leaf number in Arabidopsis through the gene *ERECTA*, and found it acts as an "ecological amplifier" (Hall et al., 2007a; Mandel et al., 2014). In this scenario, it would be possible to hypothesise that rosette shape variation could be organized across a latitude, longitude, altitude or according to ecological clues, such as living in forest or open spaces. Similarly, phenotypic correlation with flowering time or ecologically relevant traits could be studied (Pigliucci, 2002; Mitchell-Olds and Schmitt, 2006; Shindo et al., 2007; Krämer, 2015, and references therein) . It has been observed that some accessions, for example Cvi and Ag, have similar rosette development within an experiment, but relatively different between experiments. This genetics by environment interaction could be interpreted as phenotypic plasticity, since it does not seem to be a passive response, but specific to the conditions. To connect with previously mentioned observations, the phytochromes *PHYB* and *PHYD* are responsible for the shade avoidance syndrome, inducing an extension in petiole length when the leaves are under a red-infrared light ratio that indicate some other plants are shading them. It could be investigated if the distribution of *PHYB* or *PHYD* correlates with

the growing environment of each of our accessions in chapter 2, since the original location has been recorded and can be searched in mapping infrastructures such Google Earth. Transplant experiments can guide to study phenotypic plasticity and possible genes related with such plasticity (Montesinos-Navarro et al., 2010; Savolainen et al., 2013).

This thesis represent the first time, to my knowledge, that computer vision shape descriptors are applied to whole rosette shape in a genetic context. The idea of shape descriptors as objective measurements was addressed by Camargo et al. (2014). The use of accurate measurements aim to solve the difficulty in describing shape. It is common to see publications that refer to mutants in leaf or rosette shape using broad terms as "round", "loose", or simply showing pictures of control and mutants that allows the reader the interpretation of shape changes. In my experience, the selection of computer vision shape descriptors is not as explicit as intended by Camargo et al. (2014). The computation of shape descriptors is robust in terms of the same object would have similar values in several measurements even in different devices. However, two problems come up with such descriptors. On one side, they are still difficult to interpret, despite their clear names such roundness or compactness. As an example, the formula used here for roundness, $\frac{Area}{Perimeter^2}$ has been used for the compactness of object whose perimeter has concave curvatures. The situation is more complex with descriptors that have complex formulations as eccentricity. Therefore, the interpretation of shape descriptors results in confused descriptions of how the change through time, or the possible relationship of a shape descriptors with sub-rosette traits such as petiole length. On the other side, the interpretation of shape descriptors is obscured by the non-linearity of their units (Pirard and Dislaire, 2011, 2005). This means that it is difficult to interpret, for example, what is the difference between roundness value of 0.20 and 0.25. A well known problem with such descriptors is that they are not unique (Young et al., 1974), meaning that two objects with different shape has the same value, which is called in mathematics that the measurements are not "injective". This subject was treated with detail in the decade of 1980 and onwards, resulting in the development of modern ways of measuring shape like outline analysis by Fourier descriptors or the various techniques of Geometrical Morphometrics (Blum, 1973; Loncaric, 1998; Claude, 2008).

Recently, new methods to deconstruct images of Arabidopsis rosettes in their components, i.e. leaves, will facilitate a more detailed analysis of the plant architecture (Minervini et al.,

2014; Giuffrida et al., 2015; Scharr et al., 2016; Viaud et al., 2017; Minervini et al., 2017). I tried to develop a similar methodology from scratch for the first two years of this Ph.D. The intention was to study in depth the relationship of Camargo et al. (2014) shape descriptor with the segmentation quality to assess the robustness of each measurement regarding artefacts and rosette missing parts. In addition, my unsuccessful attempts of isolate leaf had the intention of correlate analytically particular leaf traits, such as blade shape, petiole length, etc. I grown several mutants for phytochrome B and D in different Red:Far-Red ratio, that actually showed a certain degree of shade avoidance syndrome. By multivariate analysis, such correspondence or redundancy analysis, the equivalence between leaf traits and rosette traits would be potentially unveiled, so the descriptions in the successive chapters results were more informative.

In my opinion, better shape descriptors could be formulated, e.g. shape context (Belongie et al., 2000). Actually recent studies are approaching the development of improved plant shape descriptions in the context of functional-structural modelling of plants, that incorporate plant genetics, physiology and shape (Letort et al., 2007; Mathieu et al., 2009; Vos et al., 2009; Bongers et al., 2014; Balduzzi et al., 2017).

Bibliography

Abràmoff, M. D., Magalhães, P. J., and Ram, S. J. Image processing with ImageJ. *Biophotonics international*, 11(7):36–42, 2004.

Adams, M. D. and Sekelsky, J. J. From sequence to phenotype: reverse genetics in Drosophila melanogaster. *Nature Reviews Genetics*, 3(3):189–198, 2002.

Alonso, J. M. and Ecker, J. R. Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. *Nature Reviews Genetics*, 7(7):524–536, 2006.

Alonso-Blanco, C., Aarts, M. G., Bentsink, L., Keurentjes, J. J., Reymond, M., Vreugdenhil, D., and Koornneef, M. What has natural variation taught us about plant development, physiology, and adaptation? *The Plant Cell*, 21(7):1877–1896, jul 2009.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R. A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T. P., Mott, R., Muliyati, N. W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P. Y., Picó, F. X., Platzer, A., Rabanal, F. A., Rodriguez, A., Rowan, B. A., Salomé, P. A., Schmid, K. J., Schmitz, R. J., Ümit Seren, Sperone, F. G., Sudkamp, M., Svardal, H., Tanzer, M. M., Todd, D., Volchenboum, S. L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., and Zhou, X. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell*, 166(2):481–491, jul 2016.

Ambrose, B. A. and Purugganan, M. Genomics, adaptation, and the evolution of plant form. In Shepard, K., editor, *The Evolution of Plant Form, Annual Plant Reviews Volume 45*, pages 189–225. John Wiley & Sons, Ltd., 2013.

Apelt, F., Breuer, D., Nikoloski, Z., Stitt, M., and Kragler, F. Phytotyping4d: a light-field imaging system for non-invasive and accurate monitoring of spatio-temporal plant growth. *The Plant Journal*, 82(4):693–706, 2015.

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P., and Nordborg, M. Genome-wide association mapping in Arabidopsis thaliana identifies previously known genes responsible for variation in flowering time and pathogen resistance. *PLoS Genetics*, preprint(2005):e60, 2005.

Arvidsson, S., Pérez-Rodríguez, P., and Mueller-Roeber, B. A growth phenotyping pipeline for Arabidopsis thaliana integrating image analysis and rosette area modeling for robust quantification of genotype effects. *New Phytologist*, 191(3):895–907, 2011.

Astle, W. and Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, nov 2009.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J., and Nordborg, M. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, mar 2010.

Bac-Molenaar, J. A., Granier, C., Keurentjes, J. J. B., and Vreugdenhil, D. Genome-wide association mapping of time-dependent growth responses to moderate drought stress inArabidopsis. *Plant, Cell & Environment*, 39(1):88–102, nov 2015.

Balasubramanian, S., Schwartz, C., Singh, A., Warthmann, N., Kim, M. C., Maloof, J. N., Loudet, O., Trainer, G. T., Dabi, T., Borevitz, J. O., et al. Qtl mapping in new Arabidopsis thaliana advanced intercross-recombinant inbred lines. *PLoS One*, 4(2):e4318, 2009.

Balding, D. J. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, oct 2006.

Balduzzi, M., Binder, B. M., Bucksch, A., Chang, C., Hong, L., Iyer-Pascuzzi, A. S., Pradal, C., and Sparks, E. E. Reshaping plant biology: Qualitative and quantitative descriptors for plant morphology. *Frontiers in Plant Science*, 08, feb 2017.

Banta, J. A., Ehrenreich, I. M., Gerard, S., Chou, L., Wilczek, A., Schmitt, J., Kover, P. X., and Purugganan, M. D. Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in Arabidopsis thaliana. *Ecology Letters*, 15(8):769–777, may 2012.

Bar, M. and Ori, N. Leaf development and morphogenesis. *Development*, 141(22):4219–4230, 2014.

Barton, M. K. and Poethig, R. S. Formation of the shoot apical meristem in Arabidopsis thaliana: an analysis of development in the wild type and in the shoot meristemless mutant. *Development*, 119(3):823–831, 1993.

Barton, N. H. and Keightley, P. D. Understanding quantitative genetic variation. *Nature Reviews Genetics*, 3(1):11–21, 2002.

Barton, N. H., Etheridge, A. M., and Véber, A. The infinitesimal model. *bioRxiv*, 2016.

Belongie, S., Malik, J., and Puzicha, J. Shape context: A new descriptor for shape matching and object recognition. In *Nips*, volume 2, 2000.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Bergelson, J. and Roux, F. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics*, 11(12):867–879, 2010.

Blum, H. Biological shape and visual science (part i). *Journal of theoretical Biology*, 38(2): 205–287, 1973.

Bongers, F. J., Evers, J. B., Anten, N. P. R., and Pierik, R. From shade avoidance responses to plant performance at vegetation level: using virtual plant modelling as a tool. *New Phytologist*, 204(2):268–272, sep 2014.

Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., and Görlach, J. Growth stage–based phenotypic analysis of Arabidopsis a model for high throughput functional genomics in plants. *The Plant Cell*, 13(7):1499–1510, 2001.

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J., and Roux, F. Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLoS Genetics*, 6(5):e1000940, may 2010.

Brachi, B., Morris, G. P., and Borevitz, J. O. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, 12(10):232, 2011.

Bradski, G. The opencv library. *Doctor Dobbs Journal*, 25(11):120–126, 2000.

Broman, K. W. Genotype probabilities at intermediate generations in the construction of recombinant inbred lines. *Genetics*, 190(2):403–412, feb 2012.

Broman, K. Review of statistical methods for qtl mapping in experimental crosses. *Lab animal*, 30(7), 2001.

Broman, K. W. and Sen, S. *A Guide to QTL Mapping with R/qtl*, volume 46. Springer, 2009.

Brosnan, T. and Sun, D.-W. Improving quality inspection of food products by computer vision—a review. *Journal of Food Engineering*, 61(1):3–16, jan 2004.

Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., Furbank, R. T., Badger, M., Pogson, B. J., and Borevitz, J. O. Traitcapture: genomic and environment modelling of plant phenomic data. *Current opinion in plant biology*, 18:73–79, 2014.

Buckler, E. and Gore, M. An Arabidopsis haplotype map takes root. *Nature Genetics*, 39(9): 1056–1057, sep 2007.

Bucksch, A. A practical introduction to skeletons for the plant sciences. *Applications in Plant Sciences*, 2(8):1400005, aug 2014.

Bush, W. S. and Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822, dec 2012. PCOMPBIOL-D-12-01453[PII] 23300413[pmid] PLoS Comput Biol.

Camargo, A., Papadopoulou, D., Spyropoulou, Z., Vlachonasios, K., Doonan, J. H., and Gay, A. P. Objective definition of rosette shape variation using a combined computer vision and data mining approach. *PLoS ONE*, 9(5):e96889, may 2014.

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. From mutations to magic: resources for gene discovery, validation and delivery in crop plants. *Current opinion in plant biology*, 11 (2):215–221, 2008.

Charlesworth, D. and Vekemans, X. How and when did Arabidopsis thaliana become highly self-fertilising. *BioEssays*, 27(5):472–476, 2005.

Chen, D., Neumann, K., Friedel, S., Kilian, B., Chen, M., Altmann, T., and Klukas, C. Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. *The Plant Cell*, 26(12):4636–4655, 2014.

Chéné, Y., Rousseau, D., Belin, É., Garbez, M., Galopin, G., and Chapeau-Blondeau, F. Shape descriptors to characterize the shoot of entire plant from multiple side views of a motorized depth sensor. *Machine Vision and Applications*, 27(4):447–461, apr 2016.

Chenu, K., Franck, N., Dauzat, J., and Lecoeur, J. Modelling the phenotypic variability of rosette architecture of Arabidopsis thaliana in several ecotypes and mutants in response to incident radiation. In *4th International Workshop on Functional–Structural Plant Models*, pages 7–11. Citeseer, 2004.

Chenu, K., Franck, N., Dauzat, J., Barczi, J.-F., Rey, H., and Lecoeur, J. Integrated responses of rosette organogenesis, morphogenesis and architecture to reduced incident light in Arabidopsis thaliana results in higher efficiency of light interception. *Functional Plant Biology*, 32(12):1123–1134, 2005.

Cheverud, J. M. Phenotypic, genetic, and environmental morphological integration in the cranium. *Evolution*, 36(3):499, may 1982.

Cheverud, J. M. and Routman, E. J. Epistasis and its contribution to genetic variance components. *Genetics*, 139(3):1455–1461, 1995.

Chitwood, D. H., Headland, L. R., Kumar, R., Peng, J., Maloof, J. N., and Sinha, N. R. The developmental trajectory of leaflet morphology in wild tomato species. *Plant physiology*, 158 (3):1230–1240, 2012.

Chitwood, D. H., Kumar, R., Headland, L. R., Ranjan, A., Covington, M. F., Ichihashi, Y., Fulop, D., Jiménez-Gómez, J. M., Peng, J., Maloof, J. N., et al. A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. *The Plant Cell*, 25(7):2465–2481, 2013a.

Chitwood, D. H., Kumar, R., Headland, L. R., Ranjan, A., Covington, M. F., Ichihashi, Y., Fulop, D., Jiménez-Gómez, J. M., Peng, J., and Maloof, J. N. A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. *The Plant Cell Online*, 25(7):2465–2481, 2013b.

Chitwood, D. H., Rundell, S. M., Li, D. Y., Woodford, Q. L., Tommy, T. Y., Lopez, J. R., Greenblatt, D., Kang, J., and Londo, J. P. Climate and developmental plasticity: interannual variability in grapevine leaf morphology. *Plant physiology*, 170(3):1480–1491, 2016.

Churchill, G. A. and Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971, 1994.

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Scholkopf, B., Nordborg, M., Ratsch, G., Ecker, J. R., and Weigel, D. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, 317(5836):338–342, jul 2007.

Claude, J. *Morphometrics with R*. Springer Science & Business Media, 2008.

Clauw, P., Coppens, F., De Beuf, K., Dhondt, S., Van Daele, T., Maleux, K., Storme, V., Clement, L., Gonzalez, N., and Inzé, D. Leaf responses to mild drought stress in natural variants of Arabidopsis. *Plant physiology*, 167(3):800–816, 2015.

Collins, A. R. Linkage disequilibrium and association mapping. In Collins, A. R., editor, *Linkage Disequilibrium and Association Mapping: Analysis and Applications*, pages 1–15. Humana Press, Totowa, NJ, 2007.

Cookson, S. J., Radziejwoski, A., and Granier, C. Cell and leaf size plasticity in Arabidopsis: what is the role of endoreduplication? *Plant, Cell & Environment*, 29(7):1273–1283, 2006.

Crow, J. F. and Kimura, M. Evolution in sexual and asexual populations. *The American Naturalist*, 99(909):439–450, 1965.

Darvasi, A. and Soller, M. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141(3):1199–1207, 1995.

De Vylder, J., Vandenbussche, F., Hu, Y., Philips, W., and Van Der Straeten, D. Rosette tracker: an open source image analysis tool for automatic quantification of genotype effects. *Plant physiology*, 160(3):1149–1159, 2012.

Deng, W., Ying, H., Helliwell, C. A., Taylor, J. M., Peacock, W. J., and Dennis, E. S. FLOWERING LOCUS c (FLC) regulates development pathways throughout the life cycle of Arabidopsis. *Proceedings of the National Academy of Sciences*, 108(16):6680–6685, apr 2011.

Devlin, B. and Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322, sep 1995.

Dhondt, S., Wuyts, N., and Inzé, D. Cell to whole-plant phenotyping: the best is yet to come. *Trends in Plant Science*, 18(8):428–439, aug 2013.

Dhondt, S., Gonzalez, N., Blomme, J., De Milde, L., Van Daele, T., Van Akoleyen, D., Storme, V., Coppens, F., TS Beemster, G., and Inzé, D. High-resolution time-resolved imaging of in vitro Arabidopsis rosette growth. *The Plant Journal*, 80(1):172–184, 2014.

Doerge, R. W. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, 3(1):43–52, 2002.

Doerge, R. W. and Churchill, G. A. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1):285–294, 1996.

Ehrenreich, I. M., Hanzawa, Y., Chou, L., Roe, J. L., Kover, P. X., and Purugganan, M. D. Candidate gene association mapping of Arabidopsis flowering time. *Genetics*, 183(1):325–335, jul 2009.

El-Lithy, M. E., Clerkx, E. J., Ruys, G. J., Koornneef, M., and Vreugdenhil, D. Quantitative trait locus analysis of growth-related traits in a new Arabidopsis recombinant inbred population. *Plant physiology*, 135(1):444–458, 2004.

Fahlgren, N., Feldman, M., Gehan, M. A., Wilson, M. S., Shyu, C., Bryant, D. W., Hill, S. T., McEntee, C. J., Warnasooriya, S. N., Kumar, I., et al. A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in setaria. *Molecular plant*, 8(10):1520–1535, 2015a.

Fahlgren, N., Gehan, M. A., and Baxter, I. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Current opinion in plant biology*, 24:93–99, 2015b.

Falconer, D. and Mackay, T. *Introduction to quantitative genetics*, volume 4. Essex: Benjamin Cummings, 1996.

Felsenstein, J. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974.

Flood, P. J., Kruijer, W., Schnabel, S. K., van der Schoor, R., Jalink, H., Snel, J. F. H., Harbinson, J., and Aarts, M. G. M. Phenomics for photosynthesis, growth and reflectance in Arabidopsis thaliana reveals circadian and long-term fluctuations in heritability. *Plant Methods*, 12(1):14, 2016.

Furbank, R. T. and Tester, M. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in plant science*, 16(12):635–644, 2011.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Ratsch, G., and Mott, R. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365):419–423, 2011. 10.1038/nature10414.

Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2): 135–145, jan 2012.

Giuffrida, M. V., Minervini, M., and Tsaftaris, S. Learning to count leaves in rosette plants. In S. A. Tsaftaris, H. S. and Pridmore, T., editors, *Proceedings of the Computer Vision Prob-*

lems in Plant Phenotyping (CVPPP), pages 1.1–1.13. British Machine Vision Association, September 2015.

Gnan, S., Priest, A., and Kover, P. X. The genetic basis of natural variation in seed size and seed number and their trade-off using Arabidopsis thaliana MAGIC lines. *Genetics*, 198(4): 1751–1758, oct 2014.

Gopal, A., Reddy, S. P., and Gayatri, V. Classification of selected medicinal plants leaf using image processing. In *Machine Vision and Image Processing (MVIP), 2012 International Conference on*, pages 5–8. IEEE, 2012.

Granier, C. and Vile, D. Phenotyping and beyond: modelling the relationships between traits. *Current opinion in plant biology*, 18:96–102, 2014.

Granier, C., Massonnet, C., Turc, O., Muller, B., Chenu, K., and Tardieu, F. Individual leaf development in Arabidopsis thaliana: a stable thermal-time-based programme. *Annals of Botany*, 89(5):595–604, 2002.

Granier, C., Aguirrezabal, L., Chenu, K., Cookson, S. J., Dauzat, M., Hamard, P., Thioux, J.-J., Rolland, G., Bouchier-Combaud, S., Lebaudy, A., et al. Phenopsis, an automated platform for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit. *New Phytologist*, 169(3):623–635, 2006.

Green, J. M., Appel, H., Rehrig, E. M., Harnsomburana, J., Chang, J.-F., Balint-Kurti, P., and Shyu, C.-R. Phenophyte: a flexible affordable method to quantify 2d phenotypes from imagery. *Plant methods*, 8(1):45, 2012.

Gupta, P. K., Rustgi, S., and Kulwal, P. L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Molecular Biology*, 57(4):461–485, mar 2005.

Haldane, J. A defense of beanbag genetics. *International Journal of Epidemiology*, 37(3): 435–442, jun 2008.

Haley, C. S. and Knott, S. A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–324, oct 1992.

Hall, M. C., Dworkin, I., Ungerer, M. C., and Purugganan, M. Genetics of microenvironmental canalization in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 104 (34):13717–13722, aug 2007a.

Hall, M. C., Dworkin, I., Ungerer, M. C., and Purugganan, M. Genetics of microenvironmental canalization in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 104 (34):13717–13722, 2007b.

Haralick, R. M. A measure for circularity of digital figures. *IEEE Transactions on Systems, Man, and Cybernetics*, 4(SMC-4):394–396, 1974.

Haralick, R. M. and Shapiro, L. G. Glossary of computer vision terms. *Pattern recognition*, 24 (1):69–93, 1991.

Harish, B., Hedge, A., Venkatesh, O., Spoorthy, D., and Sushma, D. Classification of plant leaves using morphological features and zernike moments. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1827–1831. IEEE, 2013.

Hartmann, A., Czauderna, T., Hoffmann, R., Stein, N., and Schreiber, F. Htpheno: an image analysis pipeline for high-throughput plant phenotyping. *BMC bioinformatics*, 12(1):148, 2011.

Hayes, B. Overview of statistical methods for genome-wide association studies (GWAS). *Genome-wide association studies and genomic prediction*, pages 149–169, 2013.

Hayward, A. C., Tollenaere, R., Dalton-Morgan, J., and Batley, J. Molecular marker applications in plants. In *Methods in Molecular Biology*, pages 13–27. Springer Nature, oct 2014.

Holland, J. B. Genetic architecture of complex traits in plants. *Current opinion in plant biology*, 10(2):156–161, 2007.

Hopkins, R., Schmitt, J., and Stinchcombe, J. R. A latitudinal cline and response to vernalization in leaf angle and morphology in Arabidopsis thaliana (brassicaceae). *New Phytologist*, 179(1):155–164, 2008.

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N. W., Platt, A., Sperone, F. G., Vilhjálmsson, B. J., Nordborg, M., Borevitz, J. O., and Bergelson, J. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, 44(2):212–216, jan 2012.

Houle, D., Govindaraju, D. R., and Omholt, S. Phenomics: the next challenge. *Nature reviews genetics*, 11(12):855–866, 2010.

Howell, S. H. *Molecular genetics of plant development*. Cambridge University Press, 1998.

Hu, M.-K. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187, 1962.

Huang, X., Paulo, M.-J., Boer, M., Effgen, S., Keizer, P., Koornneef, M., and van Eeuwijk, F. A. Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences*, 108(11):4488–4493, 2011.

Humphries, J. M., Bookstein, F. L., Chernoff, B., Smith, G. R., Elder, R. L., and Poss, S. G. Multivariate discrimination by shape in relation to size. *Systematic Biology*, 30(3):291–308, sep 1981.

Humplík, J. F., Lazár, D., Husičková, A., and Spíchal, L. Automated phenotyping of plant shoots using imaging methods for analysis of plant stress responses–a review. *Plant methods*, 11(1):29, 2015.

ichi Sugiyama, S. and Gotoh, M. How meristem plasticity in response to soil nutrients and light affects plant growth in four festuca grass species. *New Phytologist*, 185(3):747–758, nov 2009.

Iivarinen, J., Peura, M., Särelä, J., and Visa, A. Comparison of combined shape descriptors for irregular objects. In *BMVC*. Citeseer, 1997.

Ingvarsson, P. K. and Street, N. R. Association genetics of complex traits in plants. *New Phytologist*, 189(4):909–922, dec 2010.

Ivakov, A. and Persson, S. Plant cell shape: modulators and measurements. *Frontiers in Plant Science*, 4, 2013.

Iwata, H. and Ukai, Y. Shape: a computer program package for quantitative evaluation of biological shapes based on elliptic fourier descriptors. *Journal of Heredity*, 93(5):384–385, 2002.

Iwata, K. Placing landmarks suitably for shape analysis by optimization. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2359–2362. IEEE, 2012.

Jansen, M., Gilmer, F., Biskup, B., Nagel, K. A., Rascher, U., Fischbach, A., Briem, S., Dreissen, G., Tittmann, S., Braun, S., et al. Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via growscreen fluoro allows detection of stress tolerance in Arabidopsis thaliana and other rosette plants. *Functional Plant Biology*, 36(11):902–914, 2009.

Jansen, R. C. and Stam, P. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136(4):1447–1455, 1994.

Jung, J.-H., Domijan, M., Klose, C., Biswas, S., Ezer, D., Gao, M., Khattak, A. K., Box, M. S., Charoensawan, V., Cortijo, S., Kumar, M., Grant, A., Locke, J. C. W., Schäfer, E., Jaeger, K. E., and Wigge, P. A. Phytochromes function as thermosensors in Arabidopsis. *Science*, 354(6314):886–889, oct 2016.

Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M. I., Lin, X., et al. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.

Kendall, D. G. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984.

Keurentjes, J. J., Willems, G., van Eeuwijk, F., Nordborg, M., and Koornneef, M. A comparison of population types used for qtl mapping in Arabidopsis thaliana. *Plant Genetic Resources*, 9(02):185–188, 2011.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., and Nordborg, M. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, 39(9):1151–1155, aug 2007.

Kjemtrup, S., Boyes, D. C., Christensen, C., McCaskill, A. J., Hylton, M., and Davis, K. Growth stage-based phenotypic profiling of plants. *Plant Functional Genomics*, pages 427–441, 2003.

Klingenberg, C. P. Morphoj: an integrated software package for geometric morphometrics. *Molecular ecology resources*, 11(2):353–357, 2011.

Klukas, C., Pape, J.-M., and Entzian, A. Analysis of high-throughput plant image data with the information system iap. *Journal of Integrative Bioinformatics (JIB)*, 9(2):16–18, 2012.

Knapp, S. and Bridges, W. Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. *Genetics*, 126(3):769–777, 1990.

Knapp, S., Bridges, W., and Birkes, D. Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics*, 79(5), may 1990.

Kooke, R., Johannes, F., Wardenaar, R., Becker, F., Etcheverry, M., Colot, V., Vreugdenhil, D., and Keurentjes, J. J. Epigenetic basis of morphological variation and phenotypic plasticity in Arabidopsis thaliana. *The Plant Cell*, 27(2):337–348, 2015.

Kooke, R., Kruijer, W., Bours, R., Becker, F., Kuhn, A., van de Geest, H., Buntjer, J., Doeswijk, T., Guerra, J., Bouwmeester, H., Vreugdenhil, D., and Keurentjes, J. J. B. Genome-wide association mapping and genomic prediction elucidate the genetic architecture of morphological traits in Arabidopsis. *Plant Physiology*, 170(4):2187–2203, feb 2016.

Korte, A. and Farlow, A. The advantages and limitations of trait analysis with gwas: a review. *Plant Methods*, 9(1):29, 2013.

Kosambi, D. D. The estimation of map distances from recombination values. *Annals of Eugenics*, 12(1):172–175, 1943.

Kover, P. X. and Mott, R. Mapping the genetic basis of ecologically and evolutionarily relevant traits in Arabidopsis thaliana. *Current opinion in plant biology*, 15(2):212–217, 2012.

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. A multiparent advanced generation inter-cross to fine-map quantitative traits in ¡italic¿Arabidopsis thaliana¡/italic¿. *PLoS Genet*, 5(7):e1000551, 2009.

Kozuka, T., Horiguchi, G., Kim, G.-T., Ohgishi, M., Sakai, T., and Tsukaya, H. The different growth responses of the Arabidopsis thaliana leaf blade and the petiole during shade avoidance are regulated by photoreceptors and sugar. *Plant and Cell Physiology*, 46(1):213–223, 2005.

Krieger, J. D. A protocol for the creation of useful geometric shape metrics illustrated with a newly derived geometric measure of leaf circularity. *Applications in Plant Sciences*, 2(8): 1400009, aug 2014.

Krämer, U. Planting molecular functions in an ecological context withArabidopsis thaliana. *eLife*, 4, mar 2015.

Kuhlemeier, C. Phyllotaxis. *Trends in plant science*, 12(4):143–150, 2007.

Kwiatkowska, D. Flowering and apical meristem growth dynamics. *Journal of Experimental Botany*, 59(2):187–201, feb 2008.

Lander, E. S. and Botstein, D. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, 1989.

Lazzeroni, L. C. A chronology of fine-scale gene mapping by linkage disequilibrium. *Statistical Methods in Medical Research*, 10(1):57–76, 2001.

Leister, D., Varotto, C., Pesaresi, P., Niwergall, A., and Salamini, F. Large-scale evaluation of plant growth in Arabidopsis thaliana by non-invasive image analysis. *Plant Physiology and Biochemistry*, 37(9):671–678, sep 1999.

Letort, V., Mahe, P., Cournede, P.-H., de Reffye, P., and Courtois, B. Quantitative genetics and functional-structural plant growth models: Simulation of quantitative trait loci detection

for model parameters and application to potential yield optimization. *Annals of Botany*, 101 (8):1243–1254, aug 2007.

Leyser, O. and Day, S. *Mechanisms in plant development*. John Wiley & Sons, 2009.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., and Borevitz, J. O. Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 107(49):21199–21204, nov 2010.

Li, Z. and Sillanpää, M. J. Dynamic quantitative trait locus analysis of plant phenomic data. *Trends in Plant Science*, 20(12):822–833, dec 2015.

Lièvre, M., Wuyts, N., Cookson, S. J., Bresson, J., Dapp, M., Vasseur, F., Massonnet, C., Tisné, S., Bettembourg, M., Balsera, C., Bédiée, A., Bouvery, F., Dauzat, M., Rolland, G., Vile, D., and Granier, C. Phenotyping the kinematics of leaf development in flowering plants: recommendations and pitfalls. *Wiley Interdisciplinary Reviews: Developmental Biology*, apr 2013.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–2399, jul 2012.

Loncaric, S. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, aug 1998.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M. A., Ümit Seren, Hellmann, I., and Nordborg, M. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature Genetics*, 45(8):884–890, jun 2013.

Lynch, M., Walsh, B., et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

Mackay, T. F. C. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, 15(1):22–33, dec 2013.

Mackay, T. F. C., Stone, E. A., and Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565–577, aug 2009. 10.1038/nrg2612.

Mackay, T. F. The genetic architecture of quantitative traits. *Annual review of genetics*, 35(1): 303–339, 2001.

Mandel, T., Moreau, F., Kutsher, Y., Fletcher, J. C., Carles, C. C., and Williams, L. E. The ERECTA receptor kinase regulates Arabidopsis shoot apical meristem size, phyllotaxy and floral meristem identity. *Development*, 141(4):830–841, feb 2014.

Mardešic, S. and Segal, J. *Shape theory: the inverse system approach*, volume 26. Elsevier, 1982.

Martin, L. B., Liebl, A. L., Trotter, J. H., Richards, C. L., McCoy, K., and McCoy, M. W. Integrator networks: Illuminating the black box linking genotype and phenotype. *Integrative and Comparative Biology*, 51(4):514–527, jun 2011.

Martínez-García, J. F., Gallemí, M., Molina-Contreras, M. J., Llorente, B., Bevilaqua, M. R. R., and Quail, P. H. The shade avoidance syndrome in Arabidopsis: The antagonistic role of phytochrome a and b differentiates vegetation proximity and canopy shade. *PLoS ONE*, 9(10):e109275, oct 2014.

Massonnet, C., Vile, D., Fabre, J., Hannah, M. A., Caldana, C., Lisec, J., Beemster, G. T., Meyer, R. C., Messerli, G., Gronlund, J. T., et al. Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories. *Plant Physiology*, 152(4):2142–2157, 2010.

Mathieu, A., Cournede, P. H., Letort, V., Barthelemy, D., and de Reffye, P. A dynamic model of plant growth with interactions between development and functional mechanisms to study plant structural plasticity related to trophic competition. *Annals of Botany*, 103(8): 1173–1186, mar 2009.

MathWorks, T. Matlab r2009b. *Natick, MA*, 2009.

Minervini, M., Abdelsamea, M. M., and Tsaftaris, S. A. Image-based plant phenotyping with incremental learning and active contours. *Ecological Informatics*, 23:35–48, 2014.

Minervini, M., Giuffrida, M. V., Perata, P., and Tsaftaris, S. A. Phenotiki: An open software and hardware platform for affordable and easy image-based phenotyping of rosette-shaped plants. *The Plant Journal*, 2017.

Mishra, Y., Jänkänpää, H. J., Kiss, A. Z., Funk, C., Schröder, W. P., and Jansson, S. Arabidopsis plants grown in the field and climate chambers significantly differ in leaf morphology and photosystem components. *BMC Plant Biology*, 12(1):6, 2012.

Mitchell-Olds, T. Complex-trait analysis in plants. *Genome Biology*, 11(4):113, 2010.

Mitchell-Olds, T. and Schmitt, J. Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature*, 441(7096):947–952, jun 2006.

Montero, R. S. and Bribiesca, E. State of the art of compactness and circularity measures. *International mathematical forum*, 4(27):1305–1335, 2009.

Montesinos-Navarro, A., Wig, J., Pico, F. X., and Tonsor, S. J. Arabidopsis thaliana populations show clinal variation in a climatic gradient associated with altitude. *New Phytologist*, 189(1):282–294, sep 2010.

Moore, C. R., Johnson, L. S., Kwak, I.-Y., Livny, M., Broman, K. W., and Spalding, E. P. High-throughput computer vision introduces the time axis to a quantitative trait map of a plant growth response. *Genetics*, 195(3):1077–1086, aug 2013.

Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C., and Flint, J. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences*, 97(23):12649–12654, 2000.

Mündermann, L., Erasmus, Y., Lane, B., Coen, E., and Prusinkiewicz, P. Quantitative modeling of Arabidopsis development. *Plant physiology*, 139(2):960–968, 2005.

Mutka, A. M. and Bart, R. S. Image-based phenotyping of plant disease symptoms. *Frontiers in plant science*, 5:734, 2015.

Napp-Zinn, K. Arabidopsis thaliana. *Handbook of flowering*, pages 492–503, 1985.

Nielsen, R. Molecular signatures of natural selection. *Annual Review of Genetics*, 39(1):197–218, dec 2005.

Nixon, M. S. and Aguado, A. S. *Feature extraction & image processing for computer vision*. Academic Press, 2012.

Nordborg, M., Borevitz, J. O., Bergelson, J., Berry, C. C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J. N., Noyes, T., Oefner, P. J., Stahl, E. A., and Weigel, D. The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, 30(2):190–193, jan 2002.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., and Bergelson, J. The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, 3(7):e196, may 2005.

Novembre, J. Pritchard, stephens, and donnelly on population structure. *Genetics*, 204(2): 391–393, oct 2016.

Ogura, T. and Busch, W. From phenotypes to causal sequences: using genome wide association studies to dissect the sequence basis for variation of plant development. *Current Opinion in Plant Biology*, 23:98–108, feb 2015.

Ohta, T. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences*, 79(6):1940–1944, 1982.

O'Malley, R. C. and Ecker, J. R. Linking genotype to phenotype using the arabidopsis unimutant collection. *The Plant Journal*, 61(6):928–940, 2010.

O'Neill, C. M., Morgan, C., Kirby, J., Tschoep, H., Deng, P. X., Brennan, M., Rosas, U., Fraser, F., Hall, C., Gill, S., et al. Six new recombinant inbred populations for the study of quantitative traits in Arabidopsis thaliana. *Theoretical and Applied Genetics*, 116(5): 623–634, 2008.

Pape, J.-M. and Klukas, C. 3-d histogram-based segmentation and leaf detection for rosette plants. In *European Conference on Computer Vision*, pages 61–74. Springer, 2014.

Passardi, F., Dobias, J., Valério, L., Guimil, S., Penel, C., and Dunand, C. Morphological

and physiological traits of three major Arabidopsis thaliana accessions. *Journal of Plant Physiology*, 164(8):980–992, 2007.

Patterson, N., Price, A. L., and Reich, D. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

Pérez-Pérez, J. M., Serrano-Cartagena, J., and Micol, J. L. Genetic analysis of natural variations in the architecture of Arabidopsis thaliana vegetative leaves. *Genetics*, 162(2):893–915, 2002.

Pérez-pérez, J. M., Rubio-díaz, S., Dhondt, S., Hernández-romero, D., Sánchez-soriano, J., Beemster, G., Ponce, M. R., and Micol, J. L. Whole organ, venation and epidermal cell morphological variations are correlated in the leaves of Arabidopsis mutants. *Plant, cell & environment*, 34(12):2200–2211, sep 2011.

Peters, J. L., Cnudde, F., and Gerats, T. Forward genetics and map-based cloning approaches. *Trends in plant science*, 8(10):484–491, 2003.

Peura, M. and Iivarinen, J. Efficiency of simple shape descriptors. In *Proceedings of the third international workshop on visual form*, volume 443, page 451. Citeseer, 1997.

Pieruschka, R. and Poorter, H. Phenotyping plants: genes, phenes and machines. *Functional Plant Biology*, 39(11):813–820, 2012.

Pigliucci, M. Ecology and evolutionary biology of Arabidopsis. *The Arabidopsis Book*, 1:e0003, jan 2002.

Pigliucci, M. Genotype–phenotype mapping and the end of the 'genes as blueprint'metaphor. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1540):557–566, 2010.

Pirard, E. and Dislaire, G. Robustness of planar shape descriptors of particles. In *Proc. Xth Annual Conf. Int. Assoc. Math. Geol., Toronto*, 2005.

Pirard, E. and Dislaire, G. Sensitivity of particle size and shape parameters with respect to digitization. In *Proceedings 13 Int. Congress for Stereology*, 2011.

Platt, A., Vilhjalmsson, B. J., and Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics*, 186(3):1045–1052, sep 2010.

Poethig, R. S. *Vegetative Phase Change and Shoot Maturation in Plants*, pages 125–152. Elsevier, 2013.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, jul 2006.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, jun 2010.

Punyasena, S. W. and Smith, S. Y. Bioinformatic and biometric methods in plant morphology. *Applications in Plant Sciences*, 2(8):1400071, aug 2014.

Rahaman, M., Chen, D., Gillani, Z., Klukas, C., Chen, M., et al. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Frontiers in plant science*, 6:619, 2015.

Rahman, H., Ramanathan, V., Jagadeeshselvam, N., Ramasamy, S., Rajendran, S., Ramachandran, M., Sudheer, P. D., Chauhan, S., Natesan, S., and Muthurajan, R. Phenomics: Technologies and applications in plant and agriculture. In *PlantOmics: The Omics of Plant Science*, pages 385–411. Springer, 2015.

Reinhardt, D. Plant architecture. *EMBO Reports*, 3(9):846–851, sep 2002.

Rieseberg, L. H., Archer, M. A., and Wayne, R. K. Transgressive segregation, adaptation and speciation. *Heredity*, 83(4):363–372, oct 1999.

Rieseberg, L. H., Widmer, A., Arntz, A. M., and Burke, B. The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358(1434): 1141–1147, 2003.

Robbelen, G. Uber heterophyllie bei Arabidopsis thaliana (l.) heynh. *Ber. dtsch. bot. Ges.*, 70: 39–44, 1957.

Rohlf, F. J. The tps series of software. *Hystrix, the Italian Journal of Mammalogy*, 26(1):9–12, 2015.

Salazar-Ciudad, I. On the origins of morphological variation, canalization, robustness, and evolvability. *Integrative and comparative biology*, 47(3):390–400, 2007.

Sasaki, E., Zhang, P., Atwell, S., Meng, D., and Nordborg, M. "missing" g x e variation controls flowering time in arabidopsis thaliana. *PLOS Genetics*, 11(10):1–18, 10 2015.

Savolainen, O., Lascoux, M., and Merilä, J. Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14(11):807–820, oct 2013.

Sax, K. The association of size differences with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genetics*, 8(6):552, 1923.

Scarcelli, N., Cheverud, J. M., Schaal, B. A., and Kover, P. X. Antagonistic pleiotropic effects reduce the potential adaptive value of the frigida locus. *Proceedings of the National Academy of Sciences*, 104(43):16986–16991, 2007.

Scharr, H., Minervini, M., French, A. P., Klukas, C., Kramer, D. M., Liu, X., Luengo, I., Pape, J.-M., Polder, G., Vukadinovic, D., et al. Leaf segmentation in plant phenotyping: a collation study. *Machine vision and applications*, 27(4):585–606, 2016.

Schmitt, J., Dudley, S. A., and Pigliucci, M. Manipulative approaches to testing adaptive plasticity: phytochrome-mediated shade-avoidance responses in plants. *the american naturalist*, 154(S1):S43–S54, 1999.

Searle, S. Phenotypic, genetic and environmental correlations. *Biometrics*, 17(3):474–480, 1961.

Sen, Ś. and Churchill, G. A. A statistical framework for quantitative trait mapping. *Genetics*, 159(1):371–387, 2001.

Seren, U. Gregor mendel institute github wiki, a. `https://github.com/Gregor-Mendel-Institute/atpolydb/wiki`[Accessed: 21.02.2017].

Seren, U. The genealogy of Arabidopsis thaliana (nsf deb 0115062), b. https://github.com/Gregor-Mendel-Institute/atpolydb/wiki/The-genealogy-of-{Arabidopsis}-thaliana-(NSF-DEB-0115062) [Accessed: 21.02.2017].

Shapiro, L. and Stockman, G. C. Computer vision. 2001. *ed: Prentice Hall*, 2001.

Sheets, H. Imp-integrated morphometrics package. *Buffalo, NY: Department of Physics, Canisius College*, 2003.

Shindo, C., Bernasconi, G., and Hardtke, C. S. Natural genetic variation in Arabidopsis: Tools, traits and prospects for evolutionary ecology. *Annals of Botany*, 99(6):1043–1054, jan 2007.

Shpak, E. D. Diverse roles ofERECTAFamily genes in plant development. *Journal of Integrative Plant Biology*, 55(12):1238–1250, oct 2013.

Simon, M., Loudet, O., Durand, S., Bérard, A., Brunel, D., Sennesal, F.-X., Durand-Tardif, M., Pelletier, G., and Camilleri, C. Quantitative trait loci mapping in five new large recombinant inbred line populations of Arabidopsis thaliana genotyped with consensus single-nucleotide polymorphism markers. *Genetics*, 178(4):2253–2264, 2008.

Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, jun 2008.

Small, C. G. *The statistical theory of shape*. Springer Science & Business Media, 2012.

Smith, J. M. *The evolution of sex*. CUP Archive, 1978.

Soller, M. and Beckmann, J. Marker-based mapping of quantitative trait loci using replicated progenies. *Theoretical and Applied Genetics*, 80(2):205–208, 1990.

Somers, K. M. Allometry, isometry and shape in principal components analysis. *Systematic Zoology*, 38(2):169, jun 1989.

Sozzani, R., Busch, W., Spalding, E. P., and Benfey, P. N. Advanced imaging techniques for the study of plant growth and development. *Trends in plant science*, 19(5):304–310, 2014.

Springate, D. A., Scarcelli, N., Rowntree, J., and Kover, P. X. Correlated response in plasticity to selection for early flowering in Arabidopsis thaliana. *Journal of Evolutionary Biology*, 24 (10):2280–2288, aug 2011.

Springate, D. A. and Kover, P. X. Plant responses to elevated temperatures: a field study on phenological sensitivity and fitness responses to simulated climate warming. *Global Change Biology*, 20(2):456–465, nov 2013.

Stumpf, M. P. H. and McVean, G. A. T. Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, 4(12):959–968, dec 2003.

Sun, G., Zhu, C., Kramer, M. H., Yang, S.-S., Song, W., Piepho, H.-P., and Yu, J. Variation explained in mixed-model association mapping. *Heredity*, 105(4):333–340, feb 2010.

Sundberg, P. Shape and size-constrained principal components analysis. *Systematic Zoology*, 38(2):166, jun 1989.

Symonds, V. V., Godoy, A. V., Alconada, T., Botto, J. F., Juenger, T. E., Casal, J. J., and Lloyd, A. M. Mapping quantitative trait loci in multiple populations of Arabidopsis thaliana identifies natural allelic variation for trichome density. *Genetics*, 169(3):1649–1658, 2005.

Takuno, S., Terauchi, R., and Innan, H. The power of qtl mapping with rils. *PloS one*, 7(10): e46545, 2012.

Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y.-L., Hu, T. T., Clark, R. M., Nasrallah, J. B., Weigel, D., and Nordborg, M. The evolution of selfing in Arabidopsis thaliana. *Science*, 317(5841):1070–1072, aug 2007.

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A. E., Buckler, E. S., and Zhang, Z. GAPIT version 2: An enhanced integrated tool for genomic association and prediction. *The Plant Genome*, 9(2):0, 2016.

Teichmann, T. and Muhr, M. Shaping plant architecture. *Frontiers in Plant Science*, 6, apr 2015.

Tessmer, O. L., Jiao, Y., Cruz, J. A., Kramer, D. M., and Chen, J. Functional approach to high-throughput plant growth analysis. *BMC systems biology*, 7(6):S17, 2013.

Thoday, J. Location of polygenes. *Nature*, 191:368–370, 1961.

Thompson, D. W. et al. On growth and form. *On growth and form.*, 1942.

Tisne, S., Barbier, F., and Granier, C. The ERECTA gene controls spatial and temporal patterns of epidermal cell number and size in successive developing leaves of Arabidopsis thaliana. *Annals of Botany*, 108(1):159–168, may 2011.

Tisne, S., Serrand, Y., Bach, L., Gilbault, E., Ben Ameur, R., Balasse, H., Voisin, R., Bouchez, D., Durand-Tardif, M., Guerche, P., et al. Phenoscope: an automated large-scale phenotyping platform offering high spatial homogeneity. *The Plant Journal*, 74(3):534–544, 2013.

Torii, K. U. The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *THE PLANT CELL*, 8(4):735–746, apr 1996.

Tsukaya, H. The leaf index: Heteroblasty, natural variation, and the genetic control of polar processes of leaf expansion. *Plant and Cell Physiology*, 43(4):372–378, apr 2002.

Tsukaya, H. Leaf shape: genetic controls and environmental factors. *International Journal of Developmental Biology*, 49(5-6):547–555, 2004.

Tsukaya, H. Leaf development. *The Arabidopsis Book*, 11:e0163, jan 2013.

Tsukaya, H., Kozuka, T., and Kim, G.-T. Genetic control of petiole length in Arabidopsis thaliana. *Plant and Cell Physiology*, 43(10):1221–1228, 2002.

Utku, H. Application of the feature selection method to discriminate digitized wheat varieties. *Journal of Food Engineering*, 46(3):211–216, nov 2000.

Valdar, W., Holmes, C. C., Mott, R., and Flint, J. Mapping in structured populations by resample model averaging. *Genetics*, 182(4):1263–1277, 2009.

Van Ooijen, J. W. LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity*, 83(5):613–624, nov 1999.

van Zanten, M., Snoek, L. B., Proveniers, M. C. G., and Peeters, A. J. M. The many functions of erecta. *Trends in Plant Science*, 14(4):214–218, 2009a.

van Zanten, M., Snoek, L. B., Proveniers, M. C., and Peeters, A. J. The many functions of ERECTA. *Trends in Plant Science*, 14(4):214–218, apr 2009b.

Vanhaeren, H., Gonzalez, N., and Inzé, D. A journey through a leaf: Phenomics analysis of leaf growth in Arabidopsis thaliana. *The Arabidopsis Book*, 13:e0181, jan 2015.

VanRaden, P. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91 (11):4414–4423, nov 2008.

Viaud, G., Loudet, O., and Cournède, P.-H. Leaf segmentation and tracking in Arabidopsis thaliana combined to an organ-scale plant model for genotypic differentiation. *Frontiers in Plant Science*, 7, jan 2017.

Vilhjálmsson, B. J. and Nordborg, M. The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2, nov 2012.

Vos, J., Evers, J. B., Buck-Sorlin, G. H., Andrieu, B., Chelle, M., and de Visser, P. H. B. Functional-structural plant modelling: a new versatile tool in crop science. *Journal of Experimental Botany*, 61(8):2101–2115, dec 2009.

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G. F., van Eck, H. J., and van Eeuwijk, F. A. Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics*, 130(1):123–135, oct 2016.

Wagner, G. P. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: Evidence for a nonrandom organization of quantitative character variation. *Journal of Mathematical Biology*, 21(1):77–95, dec 1984.

Walter, A., Scharr, H., Gilmer, F., Zierer, R., Nagel, K. A., Ernst, M., Wiese, A., Virnich, O., Christ, M. M., Uhlig, B., et al. Dynamics of seedling growth acclimation towards altered light conditions can be quantified via growscreen: a setup and procedure designed for rapid optical phenotyping of different plant species. *New Phytologist*, 174(2):447–455, 2007.

Webster, M. T. and Hurst, L. D. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends in genetics*, 28(3):101–109, 2012.

Weigel, D. Natural variation in Arabidopsis. how do we find the causal genes? *PLANT PHYSIOLOGY*, 138(2):567–568, may 2005.

Weigel, D. Natural variation in Arabidopsis: From molecular genetics to ecological genomics. *PLANT PHYSIOLOGY*, 158(1):2–22, dec 2011.

Weigel, D. and Mott, R. The 1001 genomes project for Arabidopsis thaliana. *Genome Biology*, 10(5):107, 2009.

Wilson-Sánchez, D., Rubio-Díaz, S., Muñoz-Viana, R., Pérez-Pérez, J. M., Jover-Gil, S., Ponce, M. R., and Micol, J. L. Leaf phenomics: a systematic reverse genetic screen for Arabidopsis leaf mutants. *The Plant Journal*, 79(5):878–891, 2014.

Winter, W. D. The beanbag genetics controversy: Towards a synthesis of opposing views of natural selection. *Biology and Philosophy*, 12(2):149–184, 1997.

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'connell, J. R., Mangino, M., et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011.

Yang, M., Kpalma, K., and Ronsin, J. A survey of shape feature extraction techniques, 2008.

Yoshioka, Y., Iwata, H., Ohsawa, R., and Ninomiya, S. Analysis of petal shape variation of primula sieboldii by elliptic fourier descriptors and principal component analysis. *Annals of Botany*, 94(5):657–664, 2004.

Young, I. T., Walker, J. E., and Bowie, J. E. An analysis technique for biological shape. i. *Information and control*, 25(4):357–370, 1974.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, dec 2005.

Zelditch, M. L., Swiderski, D. L., and Sheets, H. D. *Geometric morphometrics for biologists: a primer*. Academic Press, 2012.

Zeng, Z.-B. Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457–1468, 1994.

Zhang, D. and Lu, G. Review of shape representation and description techniques. *Pattern recognition*, 37(1):1–19, 2004.

Zhang, W., Collins, A., Maniatis, N., Tapper, W., and Morton, N. E. Properties of linkage disequilibrium (LD) maps. *Proceedings of the National Academy of Sciences*, 99(26):17004–17007, dec 2002.

Zhang, X., Hause, R. J., and Borevitz, J. O. Natural genetic variation for growth and development revealed by high-throughput phenotyping in Arabidopsis thaliana. *G3: Genes—Genomes— Genetics*, 2(1):29–34, 2012.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360, mar 2010.

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., and Nordborg, M. An Arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1):e4, 2007.

Zheng, C., Boer, M. P., and van Eeuwijk, F. A. A general modeling framework for genome ancestral origins in multiparental populations. *Genetics*, 198(1):87–101, 2014.

Zhou, H., Muehlbauer, G., and Steffenson, B. Population structure and linkage disequilibrium in elite barley breeding germplasm from the United States. *Journal of Zhejiang University SCIENCE B*, 13(6):438–451, jun 2012.

Zou, W. and Zeng, Z.-B. Statistical methods for mapping multiple QTL. *International Journal of Plant Genomics*, 2008:1–8, 2008.

Appendix A

Digital Geometry

Image based plant phenotyping technologies use digital pictures to extract information about the plant. From visible spectra cameras, the usual commercial cameras, colour digital pictures are taken. The use of this pictures allow to study plant colour, e.g. pigmented damage, and morphological measurements. A short introduction to digital imaging with focus in digital shape descriptors is provided as support for the most technical aspects in the chapters.

A 2D colour digital image is an array, or grid, of pixels (shorthand for picture element) sorted by position and containing the values of bright intensity for, typically, three channels of additive colours (Red, Green and Blue or other colour spaces are possible). Pixels are considered squared units in a digital screen only for representation purposes. However, each pixel is just a discrete infinitesimally small point that contains information of a region in the original scene, with the consequent information loss. The latter has mathematical consequences, the points and intensity signal related operations has to be performed using discrete mathematics techniques as approximation of the common infinitesimal calculus. Therefore, most geometrical shape descriptors are discrete version of equivalent functions from analytical geometry and physics mechanical moments (Mardešic and Segal, 1982).

Yang et al. (2008) summarize the properties that shape descriptors should obey to be useful in quantifying morphological features of object for any intended purpose. Shape descriptors are estimations of morphological features of image-represented objects. From Yang et al. (2008) :

identifiability Two objects looking similar to humans observers must have similar values for a given descriptors. Equivalently, objects that looks different must have different values. This property refers to the monotony of measures, meaning that each numeric value of a

measurement should not represent more than one property of the measured object.

Translation, rotation and scale invariance In order to represent forms, shape descriptors must not change when any of these three operations are applied.

Affine invariance This property add invariance to shear operations. This property is harder to achieve and frequently is accepted that it does not happen.

Noise resistance This property concern the robustness of descriptors again noise. The source of noise may be segmentation errors, artefacts or missing parts, or an effect of digital imaging discretization, that is, for example pixelation of border that may increase or decrease the values of perimeters.

occultation invariace A desired property is that when parts of the original object are non visible by superposition of other object in the image, the feature should remain as close as possible to the value of the original.

Statistically independent When two or more descriptors are used, their values should be statistically independent. Being independent aids into a compact representation of a shape.

Reliable Objects that show the same morphological pattern should have similar features.

Peura and Iivarinen 1997 and Iivarinen et al. 1997, indicates that using a combination of shape descriptors that do not obey the previous rules, is still efficient. Their view is that combining descriptors that are simple, general and with a certain degree of correlation, provides a new perspective in shape description. The argument is that using multivariate statistics, e.g conditional distributions of variables or Self-Organizing Maps (SOM) in the original papers, is still possible to find differences among objects categories even if the descriptors properties are not ideal.

Following, each statistical shape descriptor is described by its mathematical definition, its functional meaning as rosette overall descriptor and possible advantage and drawbacks when measuring automatically from images. For more specific treatments see books on Computer vision and Shape descriptors (Nixon and Aguado, 2012; Shapiro and Stockman, 2001) and the use in biological images has been reviewed by several authors,e.g in the context of plant sciences

Dhondt et al. (2013) or plant cell microscopy Ivakov and Persson (2013). Papers providing a good context of Shape descriptors in computer vision see Yang et al. (2008); Zhang and Lu (2004). The presented descriptor are those analysed by Camargo et al. (2014). For the more comprehensive treatment of descriptors, the glossary of computer vision terms by Haralick and Shapiro (1991) is a short review, that contains most of the descriptors in this document.

Projected Rosette Area. Hereafter PRA or Area, It is calculated as plant pixels count. Its unit is pixel number, but is generally translated to squared millimetres by a conversion factor calculated from a size-known object, as tray stickers or whole tray or pots. Therefore it is a discrete measurement of Area even when expressed in continuous units (mm^2) and it smallest error is a pixel or the area of a pixel at certain resolution.

In general, PRA is an estimate of cumulative leaf area but it does not contain any form-related information by itself. It is only a measurement of size and the extent of photosynthetic surface. It can be used to estimate rosette growth rate in time lapse dynamic imaging and also to estimate shoot biomass from Area after a validation experiment (Leister et al., 1999). The estimation of Projected Rosette Area from top view images is affected by leaf superposition and occlusion; leaves overlap covering partially, sometimes total, other leaves. Thus, PRA gives back a sub-estimated photosynthetic area. Other source of Area sub-estimation is that pictures are a flat representation of the three-dimensional rosette structure, so leaves inclination, hyponasty, and tilt influence the pixel count, as also possibly lateral leaf movement does (Lièvre et al., 2013). This source of error may affect to PRA estimation in different plants if they, as an example, vary in the amount of hyponastic growth. Consequently, plant pictures must be taken always at the same time of the day, preferably when leaves are down and the movement is reduced. On the other hand, PRA is not strongly affected by small artefacts as segmentation errors, i.e. stones, green pixels corresponding to algae or moss, while they are kept small and residual. The artefacts position does not influence Area estimation. Finally, segmentation errors as missing leaves because of colour or turning, e.g. plant pixels segmented as background, may reduce the PRA values according to the number of pixels lost.

Perimeter. Perimeter refers to object's boundary length. Several methods to stablish the contour of an imaged object are available. In general, this parameter is a pixel count

from the outer limit of plant rosette.

To first compute the outer limit, two methods are available. We can consider that the inner part of the perimeter only contains pixels in the 4-neighbourhood, so pixels belonging to the perimeter are only in the North, South, East or West position of each other. The resulting perimeter has then pixels are only connected in diagonal when the shape is curved. Instead, the delimitation of the perimeter using an 8-neighbourhood force the perimeter to contain also the NW, NE, SE and SW pixels. Therefore the perimeter would not have diagonals. The former description is valid for boundary points set up by erosion of an square structuring element, other methods could yield other estimates of perimeter. As an example, the Ramer–Douglas–Peucker algorithm approximate the contour of an object to a polygon with a restricted number of vertex, which would be useful for shapes with complex borders. To make a length measurement from this concept of perimeter, the distance between pixels is set as 1 when the two pixels are in its 4-neighbourhood (North, South, West or East), and $\sqrt{2}$ when diagonal. Thus, at least to possible metrics for perimeter length.

Arabidopsis rosette perimeter should account for sum of every single leave perimeter. However, leave occlusion reduces its values. This effect is needed to be slightly explore to understand the consequences in other descriptors that use Perimeter in their formulas as Roundness. Young plants Rosettes are star-like object whose elements, leaves, extend from meristem to leaves tips, and their perimeters account for petioles, quasi-rectangular elements, and leaves, round or elongated, elliptical, elements. When leaves occlude, at least one side of both leaves disappear, reducing massively the perimeter of this pair of leaves. In the extreme case of leaves totally occluding, when one leave cover a previous leaf, or a younger leave is contained over older one cause an ever bigger loss of perimeter values. This effect is increased when leaf are able to move because of pot handling or by themselves. In addition, image resolution and aliasing generates a rough boundary that resemble the classical problem of fractal dimension, which mean that the resolution of the image strongly modify the measurement. The latter is well known as the “coastline paradox”.

Plant Region. Some morphological parameter are calculated from regions different than the

object itself. These plant regions represents the overall surface that a plant may be covering at certain stage. As an example, the pot surface could be the maximum region that the plant could cover. It has the inconvenient of having a constant diameter and then a constant area, so that it would not express any shape information. The plant region variables that are calculated for our plants are Bounding Box, Minimum Bounding Box, Minimum Boundary Circle and Convex Hull.

Rosette Bounding Box. Rosette pixels are limited by its most extreme pixels at top, bottom, left and right sides. These four pixels delimit a rectangle covering the whole rosette. From Bounding Box, its width (VRectSizeX) and height (VRectSizeY) are provided and Bounding Box Area is calculated as $BB_{Area} = BB_{Width} \times BBHeight$.

Bounding box is a classical computer vision structure that aids in limiting the object region of interest for deeper analysis, and it rectangular shape provide useful properties when applying filters and convolutions. However, it only provides information about shape in relation to object symmetry when calculating the width/height ratio. A 1:1 ratio would indicate an square, so some degree of symmetry. It is important to notice that even when the Bounding Box would be a square, nothing can be told about the symmetry in the distribution of plant pixels inside the box.

Bounding box width, height, and its area are indicative of plant size and extent but they are an overestimation or real extent. The centre of the Bounding box is easily calculated as width/2 and height/2 plus the offset. The bounding points should correspond to biggest leaves tips, although segmentation artefacts can extend bounding box or any other region far from leaves, having a strong dependence of artefacts distribution.

Rosette Minimum Bounding Box. Due to Rosette pixels distribution, the bounding box may not be the minimum one. Applying rotations to the rosette image, or calculating rosette orientation from its moments (see below), the minimum bounding box can be found. This region share bounding box properties, but it is not a convenient structure for image processing any more. However, it becomes a bit less exaggerated estimation of plant region extent and area. Only the Area of Minimum Bounding Box is provided.

Minimum Bounding Circle. It is defined as the circle that bounds all rosette pixels that have the minimum diameter. Due to geometrical reasons, only three points are often required to build this circle, with exceptions being symmetrical objects as squares or other regular polygons. In rosettes, the three most extreme leaves tips will determine this circle and its centre, being the latter displaced from the rosette centroid according to the position of this three leaves. Minimum bounding circle offer an estimation of whole rosette extent; only the diameter of the circle is provided (DiamBoundCircle).

Convex Hull. The convex hull of an image object is, informally, the smallest irregular polygon that covers the whole object surface without having any concavity, that is, all the points inside the polygon can be connected by a line that never falls out of the polygon. Another way to say it, the convex hull is similar to place a tense rope around the object. Thus, the convex hull contains all pixels from the object plus some extra area, called "lakes" and "holes". Convex hull boundaries touch the object at many points, but at concavities, the convex hull limits are straight lines up to the next intersection object-Convex hull.

The Convex Hull is mainly determined by three or four leaves tips and their surrounding pixels, and its shape depends on plant structure. Therefore, convex hull contains some relationship with plant shape although this is not straightforward to express. It is possible to say that this is the minimum convex object covering the whole rosette and reflects the region it may be occupying if. Convex Hull shape itself is influenced by the size of the petioles and the length of leaf lamina, but the disposition of leaves around the rosette, the angle between them, their movement, either lateral or tilt, will introduce variability in this shape. Convex Hull Area (ConvexArea) and Perimeter (ConvexCircumference) are provided.

Alternative values for area and perimeter can be obtained from plant regions, by using simple formulas. Bounding box Perimeter would be $BB_{Perimeter} = 2 * BB_{Width} + 2 * BB_{Height}$, and Minimum Circle Area and Perimeter can be obtained from the diameter as $MC_{Area} = \pi * \frac{MC_{Diam}}{2}^2$ and $MC_{Perimeter=2*\pi*\frac{MC_{Diam}}{2}}$. Convex Hull Area is calculated by pixel count, and Convex Hull Perimeter by the same methods than rosette perimeter.

These metrics would be overestimation of plants Area and Perimeter, but their values would be more robust against image resolution, as it was commented in the perimeter section. These values will be used for other shape descriptors that are about to be explained.

Compactness. Compactness is calculated here as the ratio between Rosette Area and Convex Hull Area. In literature Compactness can be found as"shape facto", but this term can be confounded having the formula similar to the shape descriptor we call here Roundness (see below and Montero and Bribiesca (2009)). Compactness capture part of shape information concerning the leaf density, as similar to the Leaf Area Index parameter in the field of plant eco-physiology. This is due to the fact that Convex Hull Area is always equal or higher than PRA. Many cases might then happen; seedlings and small rosettes with few and small leaves show high values (close to 1) of compactness because having few small leaves, the convex hull almost fit to the rosette. In larger and older rosettes with many mature leaves, plant leaves cover the convex hull almost completely and goes also high. However, interesting behaviours may occur. For plants whose mature leaves has long petioles, their aspect is loose and sparse, being the distance meristem to leaves tip longer than in dense ones. Thus, the Convex Hull reflects this having big gaps, especially among leaves, that reduce compactness values. The effects of petiole length, leaf lamina length and width is worthy to be deeply studied.Compactness could be calculated by using as denominator other plant regions, Bounding Box and Minimum Circle, but they would be overestimates of the same quantity. On the other hand, some descriptor that are not in use in this thesis could assess the approach of loose/dense habits in the rosettes. Some of them are mentioned just for the sake of completeness. Lacunarity, a fractal-like approach to account for heterogeneity in the object (see FracLac for ImageJ). Convexity defects, is the study of concavities between the object and the convex hull, in the rosette case the gap between leaves, features from ranging size and maximum chords from stem to convex hull borders could be meaningful. Finally, another common version of compactness is formulated as $Compactness = \frac{Perimeter}{CH_{Perimeter}}$ that accounts for the deviation of the object from it full region coverage. A last interesting descriptor for rosette habit would be the Shape Context and the Shape Matrix, consisting in calculate the distribution of an

object in concentric circles and the output is a histogram with center to outer boundary distribution of mass (Belongie et al., 2000).

Deviation from a circle. The following set of shape descriptors are actually more involved in objects' morphology than those presented so far. However, many of them were designed for different purposes than shape description. Moments are quantities describing mechanical properties of 2D or 3D objects related with their movement under certain forces. However, physical moments have been successfully applied to other applied mathematics fields, such statistics, where mean and variances are examples of moments, and digital geometrics where they help to describe the distribution of a point cloud.

Roundness. Also called in the literature compactness, stockiness and shape factor (Montero and Bribiesca, 2009) , its value is calculated as $Roundness = \frac{Area}{Perimeter^2}$, although its original formula, based in the isoperimetric inequality, $4 \cdot \pi \cdot Area \leq Perimeter^2$, that yield the isoperimetric quotient, $Q = 4 \cdot \pi \cdot \frac{Area}{Perimeter^2}$, which reach its maximum, 1, for a circle, and gets reduced along the curve get closer to a ellipse or a flat line. The constant $4 \cdot \pi$ is removed from the formula, and the inverse is calculated to create a ratio $\frac{Perimeter^2}{Area}$ that is more likely to be higher than 1.

Rosette Roundness could be interpreted as the how circular the rosette is. The deviation from a circle happens if leaves growth longer in one direction than in other. Given that Arabidopsis phyllotaxis yield new leaves at $\approx 136°$ (Viaud et al., 2017), the deviation is motivated by the growth of the two largest leaves.

According to the formula and the previously written descriptions, Perimeter is quite variable descriptor, inducing also a strong variability in roundness. Leaves total or partial occlusion generates peaks and valleys in the estimation of perimeter, specially when measuring these variables dynamically as time series.In addition, spare rosettes are star-like shape, which Perimeter is also exaggeratedly bigger than a square or a circle. Thus, roundness may not express clearly how round is the rosette, at least not equally clear for dense rosettes than is for spares ones.

Convex Hull Roundness. Convex Hulls has polygon-like shapes, and Roundness is calculated from its perimeter and area as for rosettes. Convex Hull Roundness express more informatively the overall rosette round shaped than the proper rosette

roundness. This descriptor is more robust to occlusion than rosette roundness, although is still dependent of relative position of bigger leaves. It will not be useful to find differences among overall rosette habit, e.g loose/dense, but it could be able to capture part of the asymmetries in the rosette.

Moments. The next group of shape descriptors are related with symmetry of shape by the departure of the object from a circle. The main operation is to calculate rosette moments. The first and second degree central moments, as in statistics and mechanics, are formulated as mean and variance like equations. Mean-like moment return the centroid, which is the rosette center of mass in x and y coordinates. Variance-like moment, or second degree moment, return the dispersion of the shape around the centroid. Like in statistics, two kind of moments can be calculated, variance-like, $\frac{\sum_i (X_i - \bar{X})^2}{N}$ or $\frac{\sum_i (Y_i - \bar{Y})^2}{N}$, and covariance-like $\frac{(\sum_i (X_i - \bar{X}) \cdot (Y_i - \bar{Y}))}{N}$. This "spatial" covariance value may be interpreted as the relationship between X and Y coordinates, or statistical association between two variables, and as in multivariate statistics an ellipse can be fitted like a bivariate normal distribution that conserve the same moments. In other words, covariance provides information about the shape of the probability distribution, and a similar geometrical use is available. Once the ellipse with the same covariance function is built, geometrical study on this ellipse provides information on the original shape, in our case the rosette.

Raw Moments. The general formula for raw moments in a binary image is $M_{ij} = \sum_x \sum_y (x^i \cdot y^j)$ where x and y are the coordinates of the pixel. According to this formula $M_{00} = N = Area$ and the centroid might be calculated as $\{(\bar{x}, \bar{y})\} = \{(\frac{M_{01}}{M_{00}}, \frac{M_{10}}{M_{00}})\} = \{\frac{(\sum x)}{N}, \frac{(\sum y)}{N}\}$

Central Moments. As in statistics, moments respect the centre, i.e mean, are defined as $\mu_{ij} = \sum_x \sum_y ((X - \bar{X})^i \cdot (Y - \bar{Y})^j) = \sum_x \sum_y ((X - (\frac{\sum X)}{N})^i \cdot (Y - (\frac{\sum Y)}{N})^j)$. And the order of each moment will depend on the values assigned to i and j. With this, the covariance matrix can be organized as $Cov(I) = \begin{pmatrix} \frac{\mu_{20}}{\mu_{00}} & \frac{\mu_{11}}{\mu_{00}} \\ \frac{\mu_{11}}{\mu_{00}} & \frac{\mu_{02}}{\mu_{00}} \end{pmatrix}$ and whose values are determined by the points in the image and allow to calculate the following descriptors.

Orientation. It is calculated from the moments as $Orientation = \frac{1}{2} \cdot \arctan\left(2 \cdot \frac{\mu_{11}}{\mu_{20} - \mu_{02}}\right)$.

It deduction goes beyond the requirements of this thesis. This measurement does not describe the shape of rosettes, but provides the main direction of its distribution. It is a uniformly distributed variable, because there are not any reason that plants grows preferentially towards any direction, so no analysis should indicate any significant result for it.

Minor and Major Axis Length, their Normalization and Ratio. From covariance matrix eigenvalues and eigenvectors, the orientation, eigenvector, and the length , diagonal of eigenvalue matrix, of two perpendicular axis are obtained. They are the longest axis of the ellipse, so the same that its orientation, and its perpendicular, or shorter axis of ellipse. These lengths are a measure of rosettes size, but the ratio between them is indicative of departure from a circle and a measure of rosette asymmetry. These axis length are normalized by $AxisLengthNorm = \frac{\frac{Length^2}{Area}}{16}$ before calculating the ratio.

Eccentricity. The eccentricity of an ellipse is a metric of its "elongatedness". It is calculated from the eigenvalues of covariance matrix as $Eccentricity = \sqrt{1 - \frac{\lambda_2}{\lambda_1}}$ where $\lambda_i = \frac{(\mu_2 0' + \mu_0 2')}{2} \pm \sqrt{\frac{4 \cdot \mu_{11}'^2 + (\mu_{20}' + \mu_{02}')^2}{2}}$ and $\mu_{xy}' = \frac{\mu_{xy}}{\mu_{00}}$.

Rotational Invariant moment. This moment is calculated as $RotationalMoment = \frac{\mu_{10}^2 + \mu_{01}^2}{\mu_{00}} = \frac{((\sum_x (X_i - \bar{X})^2 + \sum_y (Y_i - \bar{Y})^2)}{Area}$ and it is one of Hu's Invariant moments (Hu, 1962). In the case of the rosettes Rotational moment correlates with compactness, but it is less sensible to artefacts that actual compactness is.

These moment based measurements leverage on the spatial distribution of pixels and capture the distortion of rosettes from a circle. Their sensibility to different rosette structures, like longer or shorter petioles or rounded/elongated blades cannot be immediately extracted from their formulas. Segmentation errors may strongly influence its values as outliers would do in multivariate statistics. When the errors are not very dramatic, as small set of pixels that disappear, or little artefacts close to leaves or plant centre, descriptor values should not be very affected. However, when pieces of soil or stones, especially if big enough, may affect the covariance matrix in such a way that principal axis and other moments do not capture the essence of the rosette.

Circularity. This is another method to describe how far the shape departs from a circle similar to roundness, but in this case make use of the boundary pixels (Haralick, 1974; Montero and Bribiesca, 2009). It is calculated by extracting pixels from boundary and calculating their distance to shape centroid called radial distance. Radial distance is used to obtain mean radial distance and its standard deviation, and the ratio $Circularity = \frac{\mu_R}{\sigma_R}$. This formula is identical to the coefficient of variation in univariate statistics. This formula is more robust to the presence of artefacts than Roundness, but it is not included in all the software for computer vision based phenotyping.

Recent research in plant phenotyping has not made much use of this kind of descriptors. Rather, the selection of morphological measurement are moved to more complex metrics, reducing partially the throughput but increasing the discrimination power and accuracy. As examples, fractal and symmetry measurements are studied by Chéné et al. (2016) using depth cameras, and most shoot plant phenotyping is done on leaves (Krieger, 2014; Punyasena and Smith, 2014, and references therein).Bucksch (2014) explore the use of skeletons in the study of 2D and 3D morphology of shoots and roots by Reeb Graphs.

Appendix B

Association Mapping in a MAGIC Population

B.1 Significant Markers

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
CPT_0	chr1	1502999	2412531	1945101	MN1_1945105	4,31	33	909532	2
CPT_0	chr1	2547759	2548244	2548244	MN1_2548265	3,67	92	485	3
CPT_0	chr1	NA	13201153	13201153	SGCSNP10165	3,66	96	NA	5
CPT_0	chr1	16251782	16644538	16644538	MASC04211	3,76	79	392756	6
CPT_0	chr1	16871886	18228436	17474215	PERL0147872	4,90	16	1356550	7
CPT_0	chr1	19434966	19502363	19502363	MN1_19506032	3,87	65	67397	8
PC2_0	chr1	17179544	17474215	17474215	PERL0147872	3,39	72	294671	7
PC8_6	chr1	494205	1042428	592984	MN1_592760	3,68	0,08	548223	1
PC8_6	chr1	1189374	2211127	1844839	MN1_1844838	4,23	32	1021753	2
PC8_6	chr1	3941311	4041374	4041374	MN1_4041372	3,63	87	100063	4
RND_1	chr1	25508098	25508227	25508227	FKF1_606	4,92	54	129	13
PC3_3	chr1	20171160	20310446	20310446	MASC00497	3,76	92	139286	9
PC3_3	chr1	20762498	21378273	21137062	MASC00290	4,17	39	615775	10
SOL_0	chr1	19434966	19502363	19502363	MN1_19506032	3,68	86	67397	8
PC3_3	chr1	21904938	21941097	21941097	MN1_21944762	3,75	92	36159	12
RND_0	chr1	25508098	25508227	25508227	FKF1_606	4,40	48	129	13
RND_2	chr1	25508098	25508227	25508227	FKF1_606	5,14	24	129	13
PC3_9	chr1	28405046	28579364	28579364	MN1_28584258	3,76	95	174318	14

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
PC2_0	chr1	12892642	13201153	13201153	SGCSNP10165	3,52	55	308511	5
PC7_1	chr1	25484058	25508227	25508227	FKF1_606	6,89	8	24169	13
PC3_3	chr1	21397428	21665899	21397981	NMSNP1_21401646	3,99	56	268471	11
PC7_0	chr1	25508098	25508227	25508227	FKF1_606	4,79	57	129	13
CPT_2	chr2	16617897	17297301	16901273	NMSNP2_16908351	4,42	17	679404	18
PC1_7	chr2	10127320	11773929	11167900	MASC05927	5,24	6	1646609	18
CPT_2	chr2	9646159	16617820	11199743	MASC05920	10,18	0	6971661	18
ISO_4	chr2	11142985	11479513	11208540	ER_472	4,16	0,02	336528	18
CPT_1	chr2	9646159	17649336	11167900	MASC05927	8,98	0	8003177	18
ISO_1	chr2	9900343	10127320	10127320	MN2_10134400	3,48	98	226977	18
CPT_5	chr2	14343638	14399791	14360784	MASC05841	3,37	86	56153	18
CPT_5	chr2	14576011	14936321	14936321	MASC06034	3,58	63	360310	18
PC1_8	chr2	9583840	9900343	9752535	MASC02928	3,77	64	316503	18
CPT_1	chr2	17997173	18027069	18027069	MN2_18034146	3,58	86	29896	18
CPT_0	chr2	9583840	17649336	14936321	MASC06034	7,57	0	8065496	18
CPT_7	chr2	13979309	14036569	13996332	MN2_14003409	3,32	98	57260	18
PC2_0	chr2	16617897	16677088	16618255	HOS1_5954	3,44	64	59191	18

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
PC2_1	chr2	7724076	8436269	8141710	PHYB_2850	4,34	23	712193	17
CPT_4	chr2	9582437	17114508	11199743	MASC05920	11,24	0	7532071	18
CPT_8	chr2	11142985	12289830	11479513	MASC05372	4,19	26	1146845	18
CPT_8	chr2	12648024	12974734	12757304	MASC02492	3,46	84	326710	18
PC2_2	chr2	8725658	16480332	11199743	MASC05920	10,54	0	7754674	18
PC2_3	chr2	9582437	16480332	11199743	MASC05920	10,41	0	6897895	18
PC2_3	chr2	16858519	17114508	16901273	NMSNP2_16908351	3,55	69	255989	18
PC2_4	chr2	9583840	16480332	11199743	MASC05920	10,57	0	6896492	18
ISO_7	chr2	9646159	11773929	11167900	MASC05927	5,49	0	2127770	18
CPT_3	chr2	9582437	16613531	11199743	MASC05920	12,08	0	7031094	18
CPT_3	chr2	16618255	16677088	16677088	MN2_16684166	3,45	79	58833	18
PC2_6	chr2	9582437	15393139	11208540	ER_472	10,29	0	5810702	18
CPT_0	chr2	17757594	17785319	17785319	MN2_17792406	3,65	97	27725	18
CPT_0	chr2	17837400	17892549	17892549	MN2_17899626	3,65	98	55149	18
PC2_1	chr2	8515254	16480332	11167900	MASC05927	11,18	0	7965078	18
PC2_7	chr2	17114508	17297301	17297301	MN2_17304379	3,38	93	182793	18
PC2_7	chr2	17785319	17852943	17852943	MN2_17860020	3,46	82	67624	18

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
ISO_4	chr2	11557526	NA	NA	ATC_2596	3,68	68	NA	NA
CPT_1	chr2	17757424	17991797	17785319	MN2_17792406	3,69	64	234373	18
PC2_8	chr2	17892549	17991797	17991797	MASC02020	3,63	68	99248	18
PC9_2	chr2	12648024	14399923	13598067	MN2_13605144	5,80	5	1751899	18
ISO_6	chr2	9583840	12974734	11208540	ER_472	9,15	0	3390894	18
PC2_9	chr2	10242327	NA	11293294	MN2_11300378	5,26	2	NA	18
PC2_5	chr2	10127320	15775911	11293294	MN2_11300378	7,18	0	5648591	18
ISO_9	chr2	10242327	11479513	11167900	MASC05927	5,02	16	1237186	18
CPT_3	chr2	16830396	17044344	16901273	NMSNP2_16908351	3,66	0,06	213948	18
PRM_1	chr2	14375406	14399791	14399791	MN2_14406873	3,35	97	24385	18
PRM_2	chr2	14313833	14343638	14343638	MN2_14350717	3,38	0,09	29805	18
PC1_8	chr2	10127320	11729352	11167900	MASC05927	4,88	9	1602032	18
PC1_8	chr2	14375406	14399791	14399791	MN2_14406873	3,48	0,1	24385	18
CPT_5	chr2	10127320	12612808	11199743	MASC05920	6,75	0	2485488	18
CPT_5	chr2	13598067	14044297	13782937	MASC02512	3,57	64	446230	18
CPT_5	chr2	14313133	14313833	14313833	FES1_1177	3,37	85	700	18
PC1_9	chr2	14311833	14399923	14360784	MASC05841	3,72	46	88090	18

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
PC1_9	chr2	14690109	15393139	15393139	MASC05384	3,57	77	703030	18
CPT_6	chr2	9752535	15558433	11293294	MN2_11300378	8,01	0	5805898	18
ISO_8	chr2	9646159	11684374	11167900	MASC05927	5,92	0	2038215	18
CPT_7	chr2	13749339	13782937	13782937	MASC02512	3,34	93	33598	18
PC2_0	chr2	9646159	16613531	11199743	MASC05920	7,44	0	6967372	18
PRM_7	chr2	9583840	11817721	11167900	MASC05927	6,56	0	2233881	18
PRM_7	chr2	14313833	14399791	14399791	MN2_14406873	3,89	38	85958	18
PC6_3	chr2	256330	311255	257326	RGA_1023	3,95	31	54925	15
PC2_2	chr2	7724076	7735540	7735540	MN2_7742622	3,57	62	11464	17
PC2_2	chr2	8143031	8144073	8144073	PHYB_5215	3,45	83	1042	17
PRM_8	chr2	14044200	15558433	14399791	MN2_14406873	4,35	0,03	1514233	18
PRM_9	chr2	9646159	NA	11167900	MASC05927	5,87	1	NA	18
PRM_9	chr2	14044200	16480332	15393139	MASC05384	4,91	5	2436132	18
PRM_9	chr2	16677088	17297301	16830396	MN2_16837474	3,77	59	620213	18
CPT_9	chr2	11293294	11324386	11324386	MASC06116	3,27	0,1	31092	18
CPT_7	chr2	10127320	13289704	11293294	MN2_11300378	6,34	2	3162384	18
PC7_5	chr2	8140504	8143031	8141710	PHYB_2850	4,84	0,05	2527	17

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
PC8_2	chr2	9646159	12180334	10933252	MASC02949	4,56	11	2534175	18
ISO_1	chr2	256330	285852	257326	RGA_1023	3,67	65	29522	15
ISO_1	chr2	322691	347764	347764	MN2_347772	3,59	78	25073	16
CPT_0	chr2	17997173	18027069	18027069	MN2_18034146	3,87	65	29896	18
PC8_7	chr2	8140504	8143031	8141710	PHYB_2850	4,12	36	2527	17
PC9_0	chr2	9583840	17297301	11479513	MASC05372	11,16	0	7713461	18
PC2_7	chr2	17892549	17991797	17991797	MASC02020	3,39	93	99248	18
ISO_4	chr2	13598067	15244621	14936321	MASC06034	4,38	17	1646554	18
PC9_2	chr2	10242327	12612808	11479513	MASC05372	8,30	0	2370481	18
RND2_0	chr2	12648024	15775753	13996332	MN2_14003409	5,10	1	3127729	18
PC9_2	chr2	14576011	14794381	14794381	MN2_14801460	4,28	92	218370	18
PC9_3	chr2	8140504	8141710	8141710	PHYB_2850	4,32	76	1206	17
PC2_9	chr2	17892549	17991797	17991797	MASC02020	3,62	73	99248	18
PC9_4	chr2	11557526	13782937	12180334	MN2_12187411	6,08	1	2225411	18
PC1_6	chr2	9646159	11817721	11167900	MASC05927	5,79	0	2171562	18
RND_2	chr2	8140504	8143031	8141710	PHYB_2850	5,15	24	2527	17
RND_2	chr2	8792595	9136962	8970375	MASC02995	4,23	94	344367	18

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
PC3_1	chr2	14360784	14399923	14399791	MN2_14406873	3,86	79	39139	18
PRM_3	chr2	14375406	14399791	14399791	MN2_14406873	3,37	98	24385	18
PC1_9	chr2	7544501	7724076	7722797	FDP_733	3,68	51	179575	17
PC1_9	chr2	9249015	9900343	9900343	MASC03019	4,07	23	651328	18
PC1_9	chr2	10127320	11684374	11167900	MASC05927	4,64	8	1557054	18
PC3_3	chr2	10893143	11729352	11167900	MASC05927	5,08	5	836209	18
PC3_4	chr2	10242327	16073182	12428271	MN2_12435349	6,21	1	5830855	18
PC1_9	chr2	15487122	15558433	15558433	MN2_15565512	3,46	91	71311	18
PRM_6	chr2	14313133	14399923	14399791	MN2_14406873	3,95	41	86790	18
PRM_6	chr2	14576011	14690109	14690109	NMSNP2_14697188	3,44	92	114098	18
PRM_6	chr2	14794381	14936321	14936321	MASC06034	3,59	75	141940	18
RND_6	chr2	8140504	8143031	8141710	PHYB_2850	5,34	6	2527	17
RND_6	chr2	8792595	9136962	9136962	MASC05584	4,39	24	344367	18
PRM_7	chr2	14794381	14936321	14936321	MASC06034	3,58	78	141940	18
PRM_8	chr2	9249015	9249141	9249141	MN2_9256220	3,65	94	126	18
PRM_8	chr2	9583840	12612808	11167900	MASC05927	6,58	0	3028968	18
RND_7	chr2	9249015	9564281	9249141	MN2_9256220	4,15	55	315266	18

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
RND_7	chr2	9579737	16617820	11208540	ER_472	12,27	0	7038083	18
RND_7	chr2	16618255	18027069	17297301	MN2_17304379	5,44	13	1408814	18
RND_8	chr2	7722958	8436269	8141710	PHYB_2850	5,00	16	713311	17
PC2_4	chr2	16830396	17044344	16886507	MASC06022	3,54	74	213948	18
CPT_9	chr2	11450266	11479513	11479513	MASC05372	3,37	82	29247	18
RND_0	chr2	9752535	16677088	14942495	MN2_14949589	6,54	2	6924553	18
RND_0	chr2	16886507	16901273	16901273	NMSNP2_16908351	4,21	61	14766	18
PC2_7	chr2	9646159	16480332	11208540	ER_472	9,68	0	6834173	18
PC2_7	chr2	16613531	16614093	16614093	HOS1_1788	3,39	91	562	18
ISO_1	chr2	9582437	9752535	9646159	MN2_9653239	3,82	56	170098	18
RND_1	chr2	9646159	16677088	11199743	MASC05920	9,96	0	7030929	18
RND_1	chr2	16858519	17044344	16901273	NMSNP2_16908351	4,57	72	185825	18
PC9_1	chr2	9646159	17297301	11479513	MASC05372	11,39	0	7651142	18
PC2_8	chr2	9646159	13258513	11208540	ER_472	8,07	0	3612354	18
RND2_0	chr2	11142985	11293294	11199743	MASC05920	3,60	49	150309	18
SOL_2	chr2	14044200	14044297	14044297	SAR1_183	3,65	87	97	18
RND2_1	chr2	11143240	11167900	11167900	MASC05927	3,41	95	24660	18

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
RND2_1	chr2	12757304	14044297	13258513	MN2_13265590	4,04	28	1286993	18
PC9_3	chr2	10242327	13749339	11424765	MASC09221	8,47	0	3507012	18
RND_5	chr2	8140504	8143031	8143031	PHYB_4171	6,61	6	2527	17
PC9_5	chr2	8140504	8143031	8143031	PHYB_4171	4,93	46	2527	17
PRM_4	chr2	10559592	11684374	11167900	MASC05927	4,33	17	1124782	18
RND_9	chr2	9752535	12180334	11167900	MASC05927	6,70	0	2427799	18
PRM_2	chr2	14360784	14399923	14399791	MN2_14406873	3,61	56	39139	18
RND_2	chr2	16858519	17044344	16901273	NMSNP2_16908351	4,45	65	185825	18
PRM_5	chr2	14576011	14936321	14936321	MASC06034	3,71	45	360310	18
PRM_4	chr2	13996332	14044297	14044297	SAR1_183	3,51	73	47965	18
PRM_4	chr2	14311833	15244621	14399791	MN2_14406873	3,99	27	932788	18
PRM_5	chr2	14360784	14399791	14399791	MN2_14406873	3,63	54	39007	18
RND_5	chr2	9752535	15775753	11199743	MASC05920	8,89	0	6023218	18
PRM_6	chr2	9583840	NA	11167900	MASC05927	5,78	2	NA	18
RND_4	chr2	16617897	17114508	16901273	NMSNP2_16908351	4,86	19	496611	18
RND2_1	chr2	14343638	14936321	14399923	MN2_14407001	3,72	54	592683	18
RND_6	chr2	16677088	17297301	16901273	NMSNP2_16908351	3,84	65	620213	18

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
RND_7	chr2	8140504	8143031	8141710	PHYB_2850	4,14	55	2527	17
PRM_9	chr2	17599070	18027069	17991797	MASC02020	4,21	24	427999	18
RND_6	chr2	9249015	16345385	11167900	MASC05927	12,79	0	7096370	18
SOL_1	chr2	13289704	15558433	14399791	MN2_14406873	6,07	1	2268729	18
RND_3	chr2	16886507	16901273	16901273	NMSNP2_16908351	3,87	87	14766	18
RND_2	chr2	9582437	16480332	11199743	MASC05920	12,37	0	6897895	18
RND_3	chr2	8140504	8143031	8141710	PHYB_2850	4,78	16	2527	17
PC2_7	chr2	16886507	17044344	16901273	NMSNP2_16908351	3,41	88	157837	18
RND_9	chr2	16618255	18395070	17991797	MASC02020	5,24	7	1776815	18
RND_8	chr2	16618255	18027069	17837400	MN2_17844494	4,93	18	1408814	18
RND_4	chr2	8725658	16617820	11199743	MASC05920	15,15	0	7892162	18
RND_3	chr2	9582437	16480332	11199743	MASC05920	14,31	0	6897895	18
SOL_2	chr2	14313133	14399923	14399791	MN2_14406873	4,03	47	86790	18
RND_4	chr2	8140504	8143031	8141710	PHYB_2850	5,41	8	2527	17
RND_8	chr2	8742966	16480332	11167900	MASC05927	12,43	0	7737366	18
RND_9	chr2	15244621	16617820	16073182	MN2_16080260	4,00	43	1373199	18
PC3_0	chr3	18099537	18368658	18368658	NMSNP3_18379643	4,12	91	269121	26

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
PC6_8	chr3	2236554	2770428	2236716	MN3_2236721	3,81	52	533874	20
CPT_0	chr3	16666699	17370484	17370484	NMSNP3_17381469	3,80	0,07	703785	25
PC4_5	chr3	8693279	9546409	9025000	MASC04560	4,59	12	853130	24
PC4_5	chr3	8448459	8643716	8460291	MASC02733	3,86	57	195257	23
PC6_9	chr3	2903318	2967872	2967872	MN3_2967877	3,47	79	64554	21
PC6_8	chr3	1689577	2232927	2216916	NMSNP3_2216922	4,59	0,01	543350	19
RND2_5	chr3	18969166	19066266	18969684	NMSNP3_18980664	3,71	88	97100	27
PC4_4	chr3	22130440	22359929	22135608	MN3_22146586	3,95	57	229489	29
PC6_8	chr3	4073911	4237502	4141096	MN3_4141103	3,68	65	163591	22
PC6_7	chr3	1803899	2216916	2216916	NMSNP3_2216922	3,91	55	413017	19
PC6_8	chr3	2903318	3344784	2967872	MN3_2967877	4,09	34	441466	21
PC4_4	chr3	20798361	21978491	21978491	NMSNP3_21989468	4,45	23	1180130	28
CPT_0	chr3	NA	NA	NA	MN3_15977654	3,68	0,09	NA	NA
CPT_3	chr5	6322452	6416383	6416383	NMSNP5_6416385	3,50	72	93931	35
CPT_7	chr5	24400995	24921885	24836116	MASC01180	3,43	78	520890	41
CPT_3	chr5	6453526	6523118	6523118	NMSNP5_6523120	3,65	62	69592	35
RND_8	chr5	25539925	25547695	25547695	MASC04437	3,54	92	7770	41

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
RND_8	chr5	25796210	26103958	25946317	MN5_25963543	3,89	58	307748	41
PC2_9	chr5	25151119	26860115	25946317	MN5_25963543	6,01	0	1708996	41
CPT_7	chr5	23903662	23929164	23929164	MASC01444	3,40	81	25502	41
CPT_9	chr5	23798428	26203511	25946317	MN5_25963543	5,34	2	2405083	41
ISO_5	chr5	13597330	13832746	13597403	NMSNP5_13614633	3,54	0,09	235416	37
RND_9	chr5	24114077	26860115	25946317	MN5_25963543	5,69	1	2746038	41
CPT_8	chr5	23705451	26203511	25946317	MN5_25963543	4,79	8	2498060	41
PC4_5	chr5	14286477	14935674	14644122	NMSNP5_14661352	4,09	42	649197	38
PC2_9	chr5	23903662	24114077	23929164	MASC01444	3,62	72	210415	41
PC2_9	chr5	24357567	25150213	24979663	MN5_24996889	4,00	0,03	792646	41
CPT_4	chr5	6322452	6621146	6523118	NMSNP5_6523120	3,81	41	298694	35
CPT_2	chr5	6453526	6523118	6523118	NMSNP5_6523120	3,40	91	69592	35
CPT_8	chr5	23253768	23300895	23300895	MASC04571	3,48	79	47127	39
CPT_8	chr5	23396016	23559674	23400832	MASC01545	3,42	87	163658	40
ISO_5	chr5	12949235	13255136	13255136	NMSNP5_13272366	3,52	92	305901	36
PC2_8	chr5	25802730	26103958	25946317	MN5_25963543	3,74	0,06	301228	41
CPT_7	chr5	25802730	26012213	25946317	MN5_25963543	3,49	66	209483	41

Table B.1: Significantly associated Markers

. . . continued

phenotype	chromosome	From	To	Peak	peak.SNP	logP	gwpval	island.size	Interval
CPT_8	chr4	11488346	11579827	11579827	MN4_11579839	3,39	95	91481	31
PC9_1	chr4	11470264	11580131	11580131	MN4_11580143	4,73	85	109867	31
RND_0	chr4	15329734	15828431	15765115	NMSNP4_15765120	4,22	0,06	498697	32
PC2_4	chr4	11488346	11579827	11579827	MN4_11579839	3,42	89	91481	31
RND_8	chr4	11470264	11822736	11786400	MASC04642	3,82	67	352472	31
RND_2	chr4	14791611	15897262	NA	MN4_15087653	5,11	24	1105651	NA
RND_9	chr4	17772130	17803787	17803787	MN4_17803780	3,61	88	31657	34
RND_0	chr4	17005167	17038397	17038397	NMSNP4_17038400	4,01	83	33230	32
RND_1	chr4	17352236	17442975	17442975	MN4_17442969	4,30	94	90739	33
ECC_6	chr4	9197493	9198009	9198009	PHYD_2290	3,71	79	516	30
PC2_8	chr4	11470264	11579827	11488346	MN4_11488359	3,81	51	109563	31
RND_1	chr4	14791611	17038397	15489079	MASC03336	5,28	44	2246786	32

Table B.1: Significantly associated Markers
gwpval = Genome Wide P Value
CPT = Compactness. ISO=Isotropy
PRM = Perimeter. RND = Roundness
PCx = Principal Component x ECC = Eccentricity

Pheno	SNP	Bur	Can	Col	Ct	Edi	Hi	Kn	Ler	Mt	No	Oy	Po	Rsch	Sf	Tsu	Wil	Ws	Wu	Zu
CPT_0	MASC04211	0.65	0.64	0.62	0.66	0.69	0.65	0.64	0.68	0.66	0.64	0.65	0.64	0.65	0.64	0.64	0.66	0.68	0.65	0.64
CPT_0	MASC06034	0.61	0.72	0.66	0.64	0.63	0.63	0.63	0.71	0.67	0.65	0.67	0.68	0.69	0.65	0.64	0.64	0.65	0.63	0.67
CPT_0	MN1_1945105	0.63	0.66	0.64	0.66	0.65	0.65	0.62	0.67	0.66	0.69	0.68	0.68	0.7	0.62	0.63	0.63	0.67	0.65	0.64
CPT_0	MN1_19506032	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.62	NA	NA	NA	NA	NA	NA	NA
CPT_0	MN1_2548265	0.64	0.66	0.64	0.66	0.65	0.65	0.62	0.67	0.66	0.69	0.67	0.67	0.7	0.63	0.63	0.63	0.66	0.65	0.64
CPT_0	MN2_17792406	NA	NA	NA	NA	0.62	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CPT_0	MN2_17899626	NA	NA	NA	NA	0.63	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CPT_0	MN2_18034146	0.63	0.67	0.67	0.63	0.65	0.64	0.64	0.68	0.67	0.65	0.67	0.65	0.68	0.65	0.66	0.65	0.64	0.63	0.66
CPT_0	MN3_15977654	0.67	0.69	0.64	0.62	0.66	0.66	0.68	0.69	0.63	0.66	0.65	0.65	0.68	0.62	0.65	0.65	0.65	0.62	0.64
CPT_0	NMSNP3_17381469	0.66	0.68	0.63	0.63	0.67	0.68	0.68	0.67	0.65	0.65	0.65	0.65	0.69	0.61	0.66	0.66	0.64	0.63	0.64
CPT_0	PERL0147872	0.65	0.65	0.62	0.66	0.69	0.65	0.64	0.68	0.66	0.64	0.65	0.65	0.65	0.65	0.64	0.66	0.69	0.64	0.64
CPT_0	SGCSNP10165	NA	0.62	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CPT_1	MASC05927	0.58	0.68	0.62	0.63	0.61	0.65	0.61	0.71	0.63	0.63	0.63	0.63	0.63	0.6	0.62	0.62	0.59	0.61	0.63
CPT_1	MN2_17792406	NA	NA	NA	NA	0.6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CPT_1	MN2_18034146	0.61	0.64	0.64	0.61	0.61	0.62	0.61	0.66	0.65	0.62	0.64	0.62	0.65	0.62	0.62	0.62	0.6	0.61	0.63
CPT_2	MASC05920	0.58	0.67	0.6	0.61	0.59	0.64	0.6	0.7	0.62	0.62	0.62	0.62	0.62	0.59	0.6	0.6	0.58	0.6	0.63
CPT_2	NMSNP2_16908351	0.6	0.65	0.63	0.6	0.59	0.61	0.6	0.65	0.65	0.6	0.63	0.61	0.64	0.6	0.6	0.61	0.58	0.6	0.62
CPT_2	NMSNP5_6523120	0.6	0.64	0.58	0.57	0.61	0.63	0.61	0.62	0.64	0.59	0.59	0.63	0.63	0.61	0.63	0.61	0.61	0.65	0.64
CPT_3	MASC05920	0.59	0.67	0.61	0.6	0.59	0.65	0.61	0.71	0.61	0.63	0.62	0.63	0.62	0.6	0.62	0.61	0.58	0.6	0.64
CPT_3	MN2_16684166	0.61	0.65	0.64	0.6	0.59	0.62	0.61	0.65	0.64	0.61	0.63	0.62	0.64	0.63	0.61	0.61	0.6	0.6	0.64
CPT_3	NMSNP2_16908351	0.61	0.65	0.64	0.61	0.58	0.62	0.62	0.65	0.64	0.61	0.63	0.62	0.64	0.61	0.62	0.61	0.59	0.6	0.63
CPT_3	NMSNP5_6416385	0.6	0.62	0.59	0.58	0.61	0.64	0.61	0.62	0.64	0.6	0.6	0.64	0.63	0.62	0.63	0.61	0.63	0.67	0.64
CPT_3	NMSNP5_6523120	0.6	0.63	0.59	0.58	0.61	0.64	0.61	0.61	0.64	0.6	0.6	0.63	0.63	0.62	0.63	0.61	0.63	0.67	0.64
CPT_4	MASC05920	0.59	0.67	0.6	0.59	0.59	0.65	0.61	0.7	0.61	0.63	0.61	0.63	0.61	0.6	0.61	0.6	0.58	0.6	0.63
CPT_4	NMSNP5_6523120	0.6	0.62	0.59	0.57	0.59	0.63	0.61	0.6	0.63	0.59	0.6	0.63	0.63	0.62	0.62	0.61	0.63	0.66	0.64
CPT_5	FES1_1177	0.62	0.68	0.64	0.63	0.59	0.63	0.62	0.7	0.64	0.64	0.62	0.67	0.65	0.62	0.63	0.61	0.62	0.61	0.67
CPT_5	MASC02512	0.62	0.68	0.63	0.63	0.61	0.64	0.62	0.69	0.64	0.64	0.62	0.66	0.64	0.62	0.64	0.61	0.62	0.61	0.67
CPT_5	MASC05841	0.62	0.68	0.64	0.63	0.59	0.63	0.63	0.7	0.64	0.63	0.62	0.67	0.65	0.63	0.63	0.62	0.62	0.61	0.67
CPT_5	MASC05920	0.62	0.69	0.63	0.64	0.6	0.67	0.65	0.72	0.62	0.65	0.63	0.64	0.63	0.63	0.63	0.62	0.6	0.63	0.65
CPT_5	MASC06034	0.61	0.69	0.65	0.63	0.59	0.64	0.62	0.69	0.64	0.64	0.62	0.67	0.65	0.63	0.64	0.61	0.62	0.61	0.68
CPT_6	MN2_11300378	0.63	0.68	0.62	0.62	0.62	0.67	0.65	0.72	0.62	0.64	0.64	0.66	0.64	0.63	0.63	0.62	0.61	0.62	0.65
CPT_7	MASC01180	0.69	0.67	0.65	0.65	0.64	0.6	0.66	0.67	0.67	0.66	0.67	0.67	0.67	0.64	0.67	0.64	0.63	0.64	0.64
CPT_7	MASC01444	0.69	0.66	0.66	0.65	0.64	0.61	0.65	0.67	0.67	0.66	0.66	0.67	0.67	0.65	0.66	0.64	0.62	0.64	0.64
CPT_7	MASC02512	0.64	0.69	0.66	0.65	0.65	0.67	0.65	0.7	0.66	0.65	0.65	0.69	0.66	0.65	0.64	0.65	0.63	0.62	0.66
CPT_7	MN2_11300378	0.65	0.69	0.64	0.64	0.63	0.69	0.66	0.73	0.65	0.65	0.66	0.68	0.65	0.65	0.65	0.64	0.63	0.63	0.66
CPT_7	MN2_14003409	0.64	0.7	0.66	0.65	0.65	0.67	0.64	0.7	0.66	0.65	0.65	0.69	0.66	0.66	0.64	0.65	0.63	0.63	0.66
CPT_7	MN5_25963543	0.67	0.66	0.65	0.67	0.66	0.6	0.65	0.67	0.68	0.66	0.67	0.67	0.66	0.64	0.67	0.63	0.63	0.64	0.65
CPT_8	MASC01545	0.68	0.65	0.65	0.65	0.64	0.62	0.65	0.66	0.66	0.65	0.67	0.67	0.66	0.65	0.66	0.64	0.63	0.65	0.65

Table B.2: Parental of Origin effects

...continued

Pheno	SNP	Bur	Can	Col	Ct	Edi	Hi	Kn	Ler	Mt	No	Oy	Po	Rsch	Sf	Tsu	Wil	Ws	Wu	Zu
CPT_8	MASC02492	0.65	0.69	0.66	0.64	0.65	0.67	0.65	0.69	0.66	0.65	0.65	0.68	0.65	0.66	0.65	0.65	0.63	0.63	0.66
CPT_8	MASC04571	0.68	0.65	0.65	0.65	0.64	0.61	0.65	0.66	0.66	0.65	0.67	0.67	0.66	0.65	0.65	0.64	0.64	0.65	0.65
CPT_8	MASC05372	0.65	0.68	0.65	0.64	0.63	0.68	0.66	0.72	0.65	0.65	0.66	0.67	0.64	0.65	0.66	0.64	0.64	0.63	0.65
CPT_8	MN4_11579839	0.65	0.67	0.67	0.65	0.72	0.64	0.63	0.63	0.65	0.66	0.64	0.64	0.64	0.67	0.67	0.67	0.64	0.66	0.66
CPT_8	MN5_25963543	0.67	0.66	0.64	0.66	0.66	0.6	0.66	0.67	0.68	0.65	0.67	0.67	0.66	0.63	0.67	0.63	0.63	0.65	0.65
CPT_9	MASC05372	0.67	0.68	0.66	0.65	0.64	0.69	0.67	0.73	0.65	0.66	0.67	0.68	0.66	0.66	0.67	0.66	0.66	0.63	0.66
CPT_9	MASC06116	0.67	0.68	0.66	0.65	0.64	0.68	0.67	0.73	0.65	0.66	0.67	0.68	0.66	0.66	0.67	0.66	0.66	0.64	0.66
CPT_9	MN5_25963543	0.69	0.66	0.66	0.66	0.68	0.62	0.67	0.67	0.69	0.66	0.68	0.68	0.67	0.64	0.69	0.63	0.63	0.67	0.67
ECC_6	PHYD_2290	0.17	0.18	0.17	0.21	0.19	0.18	0.19	0.17	0.18	0.19	0.18	0.18	0.19	0.18	0.19	0.18	0.17	0.17	0.18
ISO_1	MN2_10134400	0.68	0.73	0.69	0.72	0.7	0.72	0.7	0.74	0.72	0.7	0.7	0.68	0.71	0.68	0.68	0.71	0.7	0.71	0.72
ISO_1	MN2_347772	0.69	0.73	0.7	0.71	0.71	0.7	0.69	0.73	0.71	0.7	0.71	0.69	0.7	0.69	0.71	0.69	0.68	0.71	0.72
ISO_1	MN2_9653239	0.68	0.73	0.69	0.72	0.69	0.73	0.7	0.74	0.72	0.7	0.69	0.68	0.72	0.68	0.68	0.71	0.69	0.71	0.73
ISO_1	RGA_1023	0.69	0.74	0.71	0.71	0.71	0.69	0.68	0.73	0.7	0.7	0.71	0.69	0.7	0.68	0.71	0.69	0.69	0.71	0.72
ISO_4	ATC_2596	0.68	0.71	0.68	0.65	0.66	0.73	0.67	0.74	0.68	0.69	0.69	0.69	0.68	0.64	0.66	0.68	0.64	0.67	0.71
ISO_4	ER_472	0.65	0.7	0.67	0.66	0.66	0.72	0.66	0.75	0.68	0.69	0.69	0.69	0.68	0.64	0.66	0.69	0.65	0.68	0.71
ISO_4	MASC06034	0.69	0.73	0.69	0.67	0.66	0.7	0.64	0.73	0.7	0.68	0.69	0.71	0.71	0.67	0.64	0.67	0.64	0.65	0.74
ISO_5	NMSNP5_13272366	0.67	0.75	0.64	0.64	0.63	0.65	0.64	0.69	0.69	0.63	0.69	0.66	0.65	0.71	0.64	0.65	0.67	0.67	0.64
ISO_5	NMSNP5_13614633	0.67	0.74	0.63	0.65	0.62	0.65	0.64	0.69	0.7	0.62	0.7	0.65	0.65	0.72	0.64	0.65	0.67	0.67	0.64
ISO_6	ER_472	0.61	0.73	0.61	0.67	0.63	0.75	0.66	0.75	0.63	0.67	0.64	0.7	0.64	0.6	0.65	0.63	0.65	0.67	0.66
ISO_7	MASC05927	0.61	0.71	0.62	0.62	0.62	0.74	0.64	0.73	0.66	0.67	0.64	0.69	0.62	0.61	0.64	0.64	0.62	0.67	0.65
ISO_8	MASC05927	0.57	0.66	0.58	0.62	0.61	0.69	0.63	0.73	0.65	0.63	0.63	0.67	0.59	0.6	0.61	0.58	0.6	0.64	0.64
ISO_9	MASC05927	0.59	0.64	0.6	0.61	0.6	0.69	0.6	0.72	0.64	0.61	0.6	0.68	0.57	0.58	0.62	0.59	0.59	0.6	0.63
PC1_6	MASC05927	1.34	0.12	1.07	0.71	1.23	0.23	0.76	-0.01	0.53	0.73	0.92	0.71	0.89	1.35	0.88	1.08	1.03	0.74	0.57
PC1_7	MASC05927	1.96	0.53	1.62	1.15	1.82	0.67	1.3	0.53	0.99	1.21	1.48	1.2	1.57	1.88	1.38	1.66	1.69	1.2	1.17
PC1_8	MASC02928	2.65	1.49	1.96	1.69	2.22	1.21	1.97	1.33	1.8	1.96	1.97	1.76	2.27	2.74	1.97	2.23	2.24	2.1	1.4
PC1_8	MASC05927	2.63	1.3	2.24	1.76	2.33	1.15	2	1.01	1.68	1.89	2.04	1.82	2.18	2.66	2.02	2.38	2.28	1.94	1.64
PC1_8	MN2_14406873	2.33	1.73	1.77	2.03	2.67	1.72	2.37	1.38	1.54	1.76	1.73	1.38	1.64	2.71	1.82	2.5	1.97	2.2	1.45
PC1_9	FDP_733	NA	NA	NA	NA	NA	2.34	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC1_9	MASC03019	3.06	2.08	2.41	2.02	2.89	1.95	2.6	1.75	2.33	2.43	2.65	2.32	2.68	3.18	2.53	2.61	2.95	2.81	2.05
PC1_9	MASC05384	2.91	2.37	2.4	2.59	3.01	2.29	2.6	1.92	1.91	2.53	2.26	2.19	2.22	3.4	2.34	2.85	2.89	2.8	1.99
PC1_9	MASC05841	2.79	2.55	2.31	2.48	2.96	2.26	2.8	1.76	2.08	2.35	2.23	1.82	2.27	3.39	2.36	2.95	2.81	2.91	1.96
PC1_9	MASC05927	3.08	2	2.74	2.19	2.92	1.76	2.52	1.44	2.24	2.41	2.57	2.3	2.67	3.26	2.55	2.76	3.06	2.64	2.24
PC1_9	MN2_15565512	2.88	2.49	2.4	2.59	3	2.31	2.55	1.97	1.93	2.52	2.25	2.3	2.19	3.27	2.33	2.81	2.87	2.81	2.03
PC2_0	HOS1_5954	-0.44	0.67	0.08	-0.59	-0.43	-0.62	-0.68	0.39	0.22	-0.31	0.04	-0.32	0.23	0.81	-0.68	-0.36	-0.57	-0.45	0.01
PC2_0	MASC05920	-0.98	1.07	-0.38	-0.22	-0.55	0.17	-0.7	1.17	-0.13	-0.28	-0.15	-0.58	0.11	-0.51	-0.78	-0.14	-0.82	-0.43	0.19
PC2_0	PERL0147872	-0.13	-0.36	-0.69	-0.23	0.5	-0.38	-0.39	0.28	0.04	-0.42	-0.51	-0.2	-0.2	-0.15	-0.59	-0.37	0.81	-0.25	-0.33
PC2_0	SGCSNP10165	-0.65	-0.66	-0.78	0.22	0.35	-0.51	-0.37	0.22	0.22	-0.31	-0.5	-0.29	-0.51	-0.29	-0.34	-0.36	0.54	-0.16	-0.52

Table B.2: Parental of Origin effects

...continued

Pheno	SNP	Bur	Can	Col	Ct	Edi	Hi	Kn	Ler	Mt	No	Oy	Po	Rsch	Sf	Tsu	Wil	Ws	Wu	Zu
PC2_1	MASC05927	-1.11	0.85	-0.64	-0.48	-0.74	-0.08	-0.74	1.05	-0.28	-0.35	-0.45	-0.59	-0.28	-0.53	-0.71	-0.4	-1.11	-0.4	-0.02
PC2_1	PHYB_2850	-0.91	1.11	-0.39	-0.46	-0.76	-0.51	-0.5	0.28	-0.35	-0.34	-0.13	-0.53	-0.43	-0.69	-0.59	-0.38	-0.8	-0.32	0
PC2_2	MASC05920	-0.93	0.84	-0.44	-0.52	-0.71	0.21	-0.45	1.1	-0.3	-0.17	-0.36	-0.29	-0.3	-0.6	-0.39	-0.38	-1.04	-0.31	-0.01
PC2_2	MN2_7742622	-0.85	0.47	-0.37	-0.53	-0.61	-0.18	-0.31	0.2	-0.14	-0.08	-0.03	-0.28	-0.46	-0.55	-0.36	-0.29	-0.58	-0.18	-0.09
PC2_2	PHYB_5215	-0.84	0.71	-0.37	-0.5	-0.5	-0.24	-0.39	0.3	-0.16	-0.06	-0.07	-0.28	-0.52	-0.55	-0.39	-0.36	-0.61	-0.15	-0.04
PC2_3	MASC05920	-0.35	0.87	-0.07	-0.14	-0.27	0.64	-0.11	1.43	0.11	0.4	-0.01	0.25	-0.01	-0.16	0.14	0.03	-0.58	-0.01	0.49
PC2_3	NMSNP2_16908351	-0.19	0.69	0.53	0.02	-0.17	0.24	0.18	0.56	0.41	-0.03	0.2	0.08	0.29	-0.25	0.13	-0.12	-0.43	-0.08	0.42
PC2_4	MASC05920	-0.19	0.95	0.01	-0.17	-0.23	0.62	0.02	1.5	0.05	0.46	0.13	0.35	0.12	-0.26	0	-0.02	-0.42	-0.07	0.53
PC2_4	MASC06022	0.02	0.69	0.54	0.03	-0.22	0.2	0.16	0.64	0.55	-0.1	0.34	0.18	0.44	-0.29	0.03	-0.1	-0.36	-0.09	0.6
PC2_4	MN4_11579839	0.29	0.44	0.2	0.04	1.07	-0.27	-0.2	0.06	0.21	0.55	0.02	0.03	0.04	0.22	0.5	0.38	-0.28	0.3	0.25
PC2_5	MN2_11300378	0.06	1.35	0.16	0.5	-0.26	0.81	0.35	1.8	0.19	0.46	0.06	0.38	0.14	-0.13	0.18	-0.11	-0.35	0.06	0.53
PC2_6	ER_472	-0.3	0.98	-0.25	0.06	-0.23	0.95	0.11	1.47	-0.1	0.2	-0.06	0.56	-0.02	-0.42	0	-0.2	-0.27	-0.06	0.26
PC2_7	ER_472	-0.07	0.83	0.01	-0.05	-0.28	1.03	0.21	1.63	0.05	0.2	0.08	0.68	-0.04	-0.16	0.09	0.02	-0.16	-0.1	0.23
PC2_7	HOS1_1788	-0.01	0.02	0.67	0.1	-0.25	0.64	0.3	0.72	0.51	-0.14	0.14	0.59	0.54	-0.1	0.03	-0.08	-0.16	-0.19	0.81
PC2_7	MASC02020	NA	NA	NA	NA	0.24	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC2_7	MN2_17304379	-0.02	0.36	0.64	0.38	-0.09	0.5	0.28	0.6	0.55	-0.04	0.21	0.19	0.47	-0.26	0.04	-0.08	-0.24	-0.08	0.6
PC2_7	MN2_17860020	NA	NA	NA	NA	-0.09	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC2_7	NMSNP2_16908351	0.01	0.34	0.66	0.18	-0.2	0.54	0.31	0.71	0.56	-0.11	0.21	0.23	0.51	-0.28	0.02	-0.1	-0.29	-0.08	0.69
PC2_8	ER_472	-0.27	0.46	-0.26	-0.32	-0.37	0.55	0.12	1.45	0.04	0.12	-0.05	0.58	-0.23	-0.28	0.02	-0.28	-0.19	-0.33	-0.02
PC2_8	MASC02020	NA	NA	NA	NA	-0.19	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC2_8	MN4_11488359	0.22	0.18	0.37	0.04	0.99	0.03	-0.5	-0.27	0.07	-0.08	-0.33	-0.32	-0.11	0.47	0.16	0.1	-0.29	0.16	0.4
PC2_8	MN5_25963543	0.49	-0.11	-0.23	0.11	0.24	-0.73	0.08	0.12	0.68	-0.13	0.43	0.42	0.07	-0.32	0.29	-0.39	-0.24	-0.04	-0.17
PC2_9	MASC01444	0.79	0.02	0.07	-0.18	0.28	-0.55	-0.08	0.25	0.31	-0.16	0.45	0.4	0.18	-0.25	0.31	-0.26	-0.34	0.24	-0.07
PC2_9	MASC02020	NA	NA	NA	NA	0.02	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC2_9	MN2_11300378	0.1	0.25	-0.16	-0.24	-0.4	0.66	0.11	1.43	0.06	0.19	-0.05	0.54	-0.1	-0.19	0.16	-0.19	-0.05	-0.39	0.09
PC2_9	MN5_24996889	0.85	0.23	-0.05	-0.19	0.29	-0.65	0.03	0.34	0.44	-0.1	0.44	0.46	0.15	-0.71	0.33	-0.43	-0.3	0.19	-0.12
PC2_9	MN5_25963543	0.72	-0.03	-0.08	-0.03	0.55	-0.64	0.01	0.1	0.64	-0.2	0.48	0.51	0.09	-0.5	0.48	-0.55	-0.34	0.2	-0.01
PC3_0	NMSNP3_18379643	0.6	0.87	0.65	0.56	0.42	0.83	0.51	0.3	0.35	0.71	0.72	0.7	0.99	0.23	0.67	0.44	0.73	0.06	0.66
PC3_1	MN2_14406873	-0.45	0.23	0	-0.33	-0.17	-0.09	-0.09	0.44	0.39	0.21	0.31	0.49	0.28	-0.07	-0.02	-0.03	0.21	-0.26	-0.01
PC3_3	MASC00290	-0.46	-0.5	-0.87	-0.47	-0.23	-0.55	-0.4	-0.47	-0.81	-0.59	-0.6	-0.58	-0.55	-0.65	-0.85	-0.52	-0.39	-0.8	-0.89
PC3_3	MASC00497	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.59	NA	NA	NA	NA	NA	NA	NA
PC3_3	MASC05927	-0.86	-0.18	-0.72	-0.64	-0.8	-0.45	-0.69	0.1	-0.68	-0.61	-0.47	-0.55	-0.45	-0.67	-0.69	-0.74	-0.69	-0.8	-0.62
PC3_3	MN1_21944762	-0.56	-0.52	-0.81	-0.48	-0.22	-0.5	-0.41	-0.49	-0.85	-0.68	-0.67	-0.55	-0.59	-0.67	-0.73	-0.54	-0.38	-0.74	-0.78
PC3_3	NMSNP1_21401646	-0.46	-0.5	-0.87	-0.48	-0.25	-0.52	-0.42	-0.49	-0.81	-0.57	-0.6	-0.56	-0.56	-0.66	-0.81	-0.53	-0.4	-0.79	-0.83
PC3_4	MN2_12435349	-0.88	-0.32	-0.82	-0.92	-0.91	-0.7	-0.85	-0.3	-0.78	-0.8	-0.76	-0.66	-0.73	-0.71	-0.75	-0.89	-0.91	-0.84	-0.74
PC3_9	MN1_28584258	1.19	0.49	0.56	0.94	0.37	0.65	0.5	0.71	0.52	0.82	0.95	0.63	0.58	0.84	0.94	0.87	1.09	0.72	1.13
PC4_4	MN3_22146586	0.13	0.32	0.14	0.34	0.1	0.47	0.12	0.16	0.23	0.29	0.09	0.11	0.08	0.09	0.28	0.08	0.42	0.07	0.39

Table B.2: Parental of Origin effects

...continued

Pheno	SNP	Bur	Can	Col	Ct	Edi	Hi	Kn	Ler	Mt	No	Oy	Po	Rsch	Sf	Tsu	Wil	Ws	Wu	Zu
PC4_4	NMSNP3_21989468	0.14	0.36	0.14	0.34	0.1	0.47	0.12	0.15	0.23	0.32	0.1	0.11	0.07	0.08	0.27	0.08	0.42	0.07	0.36
PC4_5	MASC02733	0.07	0.25	0.38	0.14	0.38	0.31	0.53	-0.01	0.08	0.29	0.32	0.35	0.32	0.15	0.18	-0.19	0.18	0.3	0.1
PC4_5	MASC04560	0.01	0.25	0.39	0.15	0.4	0.23	0.54	-0.01	0.11	0.32	0.29	0.29	0.34	0.18	0.21	-0.12	0.18	0.28	0.03
PC4_5	NMSNP5_14661352	0.31	-0.19	0.38	0.24	0.3	0.28	0.29	-0.01	0.13	0.52	-0.09	0.23	0.34	0.1	0.43	0.2	0.12	0.11	0.44
PC6_3	RGA_1023	0.02	0.19	0.08	0.03	-0.02	0.08	0.08	0.17	0.09	0.02	0.09	0.17	0.05	-0.04	0.04	0.06	0.08	0.07	0.02
PC6_7	NMSNP3_2216922	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.21	NA	NA	NA	NA	NA	NA	NA
PC6_8	MN3_2236721	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.2	NA	NA	NA	NA	NA	NA	NA
PC6_8	MN3_2967877	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.17	NA	NA	NA	NA	NA	NA	NA
PC6_8	MN3_4141103	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.11	NA	NA	NA	NA	NA	NA	NA
PC6_8	NMSNP3_2216922	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.31	NA	NA	NA	NA	NA	NA	NA
PC6_9	MN3_2967877	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.24	NA	NA	NA	NA	NA	NA	NA
PC7_0	FKF1_606	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC7_1	FKF1_606	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC7_5	PHYB_2850	-0.13	-0.06	-0.13	-0.08	-0.04	-0.13	-0.2	-0.06	-0.08	-0.04	-0.06	-0.13	-0.24	-0.15	-0.08	-0.12	-0.03	-0.11	-0.07
PC8_2	MASC02949	-0.06	0.13	-0.01	-0.09	-0.01	-0.01	-0.09	-0.03	-0.03	0	-0.05	-0.05	0.01	0.05	-0.07	-0.05	-0.04	-0.04	-0.03
PC8_6	MN1_1844838	0	-0.01	-0.05	0	-0.02	-0.07	0	-0.08	-0.06	-0.06	-0.07	-0.07	0.04	0.04	0.04	0.05	-0.05	0.05	0.07
PC8_6	MN1_4041372	0.01	0	-0.04	-0.01	-0.01	-0.1	-0.04	-0.11	0	0	-0.05	-0.04	0.03	0.02	0.07	0.04	-0.07	0.01	0.03
PC8_6	MN1_592760	-0.01	-0.01	-0.04	-0.01	-0.01	-0.04	0.01	-0.09	-0.04	-0.06	-0.05	-0.06	0.02	0.04	0.02	0	-0.04	0.06	0.05
PC8_7	PHYB_2850	0.01	-0.07	0.04	-0.06	0.01	-0.05	-0.08	0.01	-0.04	0.07	-0.06	-0.02	0.01	0.05	0	-0.05	0.11	0.02	-0.03
PC9_0	MASC05372	-0.09	0.03	-0.03	-0.02	-0.1	0	-0.04	0.13	-0.04	-0.03	-0.03	-0.04	-0.01	-0.06	-0.05	-0.06	-0.07	-0.06	-0.04
PC9_1	MASC05372	-0.08	0.02	-0.05	-0.04	-0.1	-0.02	-0.06	0.09	-0.05	-0.06	-0.04	-0.04	-0.03	-0.06	-0.07	-0.07	-0.07	-0.08	-0.07
PC9_1	MN4_11580143	-0.05	-0.04	-0.04	-0.05	0.03	-0.06	-0.09	-0.05	-0.06	-0.02	-0.06	-0.05	-0.03	-0.04	-0.04	-0.07	-0.05	-0.05	-0.06
PC9_2	MASC05372	-0.05	0.02	-0.06	-0.02	-0.08	-0.01	-0.04	0.1	-0.04	-0.04	-0.01	-0.02	-0.03	-0.06	-0.04	-0.05	-0.05	-0.06	-0.05
PC9_2	MN2_13605144	-0.05	0.04	-0.04	-0.04	-0.07	-0.03	-0.06	0.06	-0.03	-0.03	-0.02	0.01	0	-0.05	-0.04	-0.06	-0.04	-0.07	-0.05
PC9_2	MN2_14801460	-0.05	0.05	-0.04	-0.03	-0.09	-0.03	-0.04	0.04	-0.03	-0.05	-0.02	0.01	0	-0.03	-0.04	-0.06	-0.05	-0.05	-0.04
PC9_3	MASC09221	-0.04	0.05	-0.02	-0.01	-0.04	0.02	-0.02	0.11	-0.02	0	-0.01	0	0.01	-0.02	-0.02	-0.03	-0.02	-0.04	-0.01
PC9_3	PHYB_2850	-0.03	0.03	-0.02	-0.01	-0.03	-0.03	-0.02	0.03	-0.01	0.02	0.03	-0.01	-0.01	-0.01	-0.02	-0.02	0	-0.02	-0.02
PC9_4	MN2_12187411	-0.01	0.07	-0.01	-0.01	0	0.03	0	0.09	0.01	0	0.01	0	0.01	-0.01	-0.02	0	0	-0.02	0.01
PC9_5	PHYB_4171	0.03	0.09	0.04	0.03	0.06	0.03	0.04	0.08	0.05	0.04	0.07	0.02	0.03	0.05	0.03	0.05	0.04	0.02	0.02
PRM_1	MN2_14406873	112.57	78.86	97.52	111.88	105.44	104.52	109.85	80.57	83.61	90.82	88.63	88.97	92.1	105.71	98.3	98.22	98.93	105.81	84.91
PRM_2	MN2_14350717	130.22	95.9	118.26	132.55	131.13	126.07	134.61	98.29	103.71	112.11	111.72	106.74	114.04	132.89	122.18	119.39	123.29	129.4	103.69
PRM_2	MN2_14406873	130.39	94	119.3	133.15	130.64	126.55	134.41	97.94	102.67	112.41	110.81	106.68	113.85	131.03	121.04	120.51	123.47	129.85	101.59
PRM_3	MN2_14406873	155.12	110.36	139.77	155.94	156.95	148.42	158.57	117.4	125.28	132.44	135.13	128.43	134.08	153.16	144.53	146.82	147.73	153.7	119.61
PRM_4	MASC05927	193.77	140.01	181.18	172.41	191.26	155.28	174.42	131.62	163.15	164.04	174.36	172.19	173.08	193.69	173.78	174.18	195.64	174.44	149.44
PRM_4	MN2_14406873	186.64	133.28	167.15	186.75	194.85	178.28	191.54	141.1	151.81	160.57	165.44	151.03	163.82	187.72	176	178.29	179.06	184.76	142.07
PRM_4	SAR1_183	183.82	139.91	166.01	185.75	188.94	176.16	191.39	142.17	153.58	160.74	169.25	158.82	167.45	186.81	175.32	177.66	178.05	177.92	151.19
PRM_5	MASC06034	223.39	165.36	197.04	211.24	239.16	229.62	224.17	162.47	172.89	197.67	197.74	194.78	194.39	232.31	199.28	221.9	210.42	219.7	156.64

Table B.2: Parental of Origin effects

...continued

Pheno	SNP	Bur	Can	Col	Ct	Edi	Hi	Kn	Ler	Mt	No	Oy	Po	Rsch	Sf	Tsu	Wil	Ws	Wu	Zu
PRM_5	MN2_14406873	224.93	171.34	194.61	209.27	238.76	223.47	222.36	155.45	174.36	197.94	197.3	177.21	201.28	232.33	204.32	221.77	213.91	224.09	153.95
PRM_6	MASC05927	271.7	191.56	255.31	234.65	268.36	198.67	235.91	181.37	222.92	237.18	248.59	230.89	241.97	278.56	235.17	251.58	260.22	236.74	212.38
PRM_6	MASC06034	260.92	194.86	230.06	250.68	278.67	239.22	260.78	204.38	205.84	233.46	234.11	218.9	215.48	270.16	237.56	257.83	242.09	244.97	202.94
PRM_6	MN2_14406873	257.13	201.28	229.74	250.09	282.45	233.59	260.21	198.27	207.95	228.64	232.29	204.18	222.22	270.14	245.48	261.37	243.07	250.64	199.02
PRM_6	NMSNP2_14697188	258.62	201.1	229.6	249	276.15	235.65	259.29	204.61	209.38	231.61	233.47	216.2	220.71	268.2	238.27	259.56	242.12	245.08	201.78
PRM_7	MASC05927	314.48	220.47	292.12	271.37	314.4	223.78	264.26	207.27	262.08	277.76	289.75	260.41	286.82	305.13	271.74	290.45	304.42	280.45	254.92
PRM_7	MASC06034	301.83	244.33	263.25	281.58	330.81	261.02	296.26	237.67	241.11	270.9	271.17	244.77	245.77	302.66	276.38	306.29	285.41	289.38	238.77
PRM_7	MN2_14406873	301.78	247.48	262.75	279.72	334.48	257.8	293.17	230.03	246.14	264.17	266.69	230.85	257.08	302.69	285.41	309.04	283.76	293.63	238.94
PRM_8	MASC05927	357.13	267.7	338.2	326.21	354.78	267.49	311.74	241.16	305.78	314	329.83	301.05	334.43	359.41	315.73	344.4	347.12	334.4	297.69
PRM_8	MN2_14406873	344.69	278.24	304.46	328.89	368.32	306.54	344.04	269.84	290.21	311.81	312.34	270.9	296.31	360.75	329.53	356.51	331.76	340.79	277.32
PRM_8	MN2_9256220	352.78	271.85	320.08	318	350.16	262.92	315.71	281.32	303.81	322.18	320.73	304.66	337.61	365.99	317.19	328.07	333.95	340.69	284.92
PRM_9	MASC02020	NA	NA	NA	NA	365.27	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PRM_9	MASC05384	388.96	318.15	342.3	360.23	389.04	342.22	352.24	312.6	314.73	355.25	342.61	329.5	328.28	396.96	345.57	386.53	389.27	382.68	308.02
PRM_9	MASC05927	382.44	309.96	381.35	343.78	392.51	304.38	347.02	271.27	342.6	346.58	361.65	343.42	360.57	392.36	348.39	368.53	390	376.78	333.39
PRM_9	MN2_16837474	385.83	334.12	341.27	367.55	396.69	331.03	349.22	320.16	295.48	361.93	334.72	376.7	335.5	397.64	344.26	380.4	385.35	381.1	312.06
RND_0	FKF1_606	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RND_0	MN2_14949589	0.2	0.29	0.22	0.2	0.2	0.21	0.21	0.26	0.24	0.22	0.23	0.22	0.25	0.22	0.23	0.21	0.21	0.21	0.23
RND_0	NMSNP2_16908351	0.2	0.27	0.23	0.2	0.2	0.21	0.21	0.25	0.25	0.22	0.24	0.21	0.24	0.22	0.23	0.21	0.21	0.22	0.22
RND_0	NMSNP4_15765120	0.24	0.25	0.19	0.23	0.25	0.21	0.22	0.25	0.22	0.27	0.22	0.22	0.23	0.21	0.23	0.21	0.2	0.23	0.2
RND_0	NMSNP4_17038400	0.21	0.24	0.2	0.23	0.24	0.21	0.22	0.24	0.22	0.26	0.22	0.22	0.24	0.21	0.23	0.22	0.2	0.23	0.21
RND_1	FKF1_606	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RND_1	MASC03336	0.2	0.2	0.16	0.2	0.22	0.18	0.19	0.2	0.19	0.24	0.19	0.19	0.19	0.18	0.2	0.18	0.17	0.19	0.17
RND_1	MASC05920	0.16	0.25	0.18	0.19	0.16	0.2	0.17	0.26	0.2	0.2	0.19	0.19	0.19	0.18	0.18	0.18	0.16	0.18	0.19
RND_1	MN4_17442969	0.16	0.2	0.17	0.21	0.21	0.18	0.19	0.2	0.18	0.23	0.19	0.19	0.2	0.18	0.19	0.18	0.18	0.19	0.17
RND_1	NMSNP2_16908351	0.18	0.22	0.2	0.17	0.17	0.19	0.18	0.22	0.22	0.18	0.2	0.18	0.2	0.18	0.18	0.19	0.17	0.18	0.19
RND2_0	MASC05920	0.84	0.88	0.85	0.85	0.85	0.86	0.84	0.87	0.85	0.85	0.85	0.84	0.86	0.85	0.83	0.86	0.84	0.85	0.86
RND2_0	MN2_14003409	0.86	0.89	0.85	0.85	0.86	0.84	0.83	0.87	0.85	0.85	0.85	0.85	0.86	0.86	0.83	0.85	0.84	0.84	0.86
RND2_1	MASC05927	0.84	0.88	0.85	0.85	0.85	0.86	0.85	0.87	0.86	0.86	0.85	0.85	0.86	0.86	0.85	0.86	0.84	0.86	0.87
RND2_1	MN2_13265590	0.86	0.89	0.86	0.86	0.86	0.85	0.85	0.87	0.86	0.85	0.85	0.85	0.86	0.86	0.85	0.85	0.84	0.86	0.86
RND2_1	MN2_14407001	0.86	0.88	0.86	0.86	0.86	0.85	0.85	0.87	0.85	0.85	0.85	0.85	0.86	0.86	0.84	0.85	0.84	0.86	0.86
RND2_5	NMSNP3_18980664	0.91	0.9	0.89	0.9	0.9	0.89	0.9	0.89	0.9	0.89	0.89	0.89	0.89	0.89	0.9	0.9	0.89	0.9	0.9
RND_2	FKF1_606	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RND_2	MASC02995	NA	NA	NA	NA	NA	0.18	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RND_2	MASC05920	0.15	0.23	0.16	0.17	0.15	0.19	0.16	0.25	0.18	0.19	0.17	0.17	0.17	0.16	0.17	0.17	0.15	0.16	0.18
RND_2	MN4_15087653	0.2	0.19	0.15	0.2	0.21	0.17	0.17	0.18	0.17	0.22	0.17	0.17	0.17	0.16	0.19	0.16	0.15	0.17	0.16
RND_2	NMSNP2_16908351	0.16	0.21	0.18	0.16	0.15	0.18	0.18	0.2	0.2	0.16	0.18	0.16	0.19	0.16	0.17	0.17	0.15	0.16	0.17
RND_2	PHYB_2850	0.16	0.24	0.17	0.17	0.15	0.17	0.17	0.21	0.17	0.18	0.2	0.18	0.17	0.16	0.16	0.17	0.16	0.17	0.18

Table B.2: Parental of Origin effects

...continued

Pheno	SNP	Bur	Can	Col	Ct	Edi	Hi	Kn	Ler	Mt	No	Oy	Po	Rsch	Sf	Tsu	Wil	Ws	Wu	Zu
RND_3	MASC05920	0.14	0.22	0.16	0.16	0.15	0.19	0.16	0.25	0.17	0.18	0.17	0.17	0.17	0.15	0.17	0.16	0.14	0.16	0.18
RND_3	NMSNP2_16908351	0.15	0.2	0.18	0.15	0.15	0.17	0.17	0.19	0.19	0.16	0.17	0.16	0.18	0.15	0.17	0.16	0.15	0.16	0.18
RND_3	PHYB_2850	0.16	0.22	0.16	0.16	0.15	0.17	0.17	0.19	0.17	0.18	0.19	0.17	0.16	0.15	0.16	0.16	0.16	0.16	0.18
RND_4	MASC05920	0.13	0.21	0.14	0.14	0.13	0.17	0.15	0.23	0.16	0.17	0.15	0.16	0.15	0.14	0.15	0.14	0.13	0.14	0.17
RND_4	NMSNP2_16908351	0.14	0.19	0.17	0.14	0.13	0.15	0.16	0.18	0.18	0.15	0.16	0.14	0.16	0.14	0.15	0.14	0.14	0.14	0.16
RND_4	PHYB_2850	0.15	0.21	0.15	0.15	0.13	0.14	0.15	0.18	0.16	0.16	0.17	0.15	0.15	0.14	0.15	0.15	0.15	0.15	0.17
RND_5	MASC05920	0.13	0.21	0.14	0.15	0.13	0.17	0.15	0.24	0.16	0.16	0.14	0.15	0.14	0.13	0.15	0.13	0.13	0.15	0.18
RND_5	PHYB_4171	0.14	0.19	0.16	0.16	0.14	0.14	0.14	0.18	0.16	0.16	0.17	0.14	0.13	0.12	0.16	0.14	0.15	0.15	0.17
RND_6	MASC05584	NA	NA	NA	NA	NA	0.14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RND_6	MASC05927	0.13	0.19	0.13	0.15	0.13	0.18	0.14	0.21	0.15	0.15	0.14	0.15	0.14	0.12	0.14	0.13	0.13	0.14	0.15
RND_6	NMSNP2_16908351	0.14	0.17	0.16	0.15	0.13	0.16	0.14	0.17	0.17	0.13	0.15	0.14	0.15	0.12	0.15	0.13	0.13	0.14	0.15
RND_6	PHYB_2850	0.13	0.2	0.15	0.14	0.13	0.14	0.15	0.17	0.15	0.15	0.17	0.15	0.13	0.12	0.14	0.14	0.14	0.14	0.16
RND_7	ER_472	0.12	0.18	0.13	0.15	0.12	0.17	0.14	0.21	0.14	0.15	0.14	0.16	0.13	0.12	0.14	0.13	0.13	0.13	0.14
RND_7	MN2_17304379	0.13	0.16	0.16	0.16	0.12	0.16	0.15	0.16	0.17	0.13	0.14	0.14	0.14	0.12	0.14	0.13	0.12	0.13	0.15
RND_7	MN2_9256220	0.12	0.18	0.14	0.15	0.13	0.17	0.15	0.18	0.15	0.15	0.14	0.15	0.13	0.12	0.14	0.13	0.13	0.14	0.15
RND_7	PHYB_2850	0.13	0.18	0.14	0.14	0.13	0.15	0.15	0.17	0.15	0.15	0.16	0.15	0.13	0.12	0.14	0.14	0.14	0.14	0.15
RND_8	MASC04437	0.15	0.13	0.13	0.13	0.13	0.11	0.13	0.15	0.15	0.13	0.14	0.14	0.13	0.12	0.15	0.12	0.12	0.14	0.12
RND_8	MASC04642	0.13	0.14	0.14	0.13	0.17	0.13	0.12	0.12	0.13	0.13	0.13	0.13	0.13	0.15	0.14	0.13	0.12	0.13	0.14
RND_8	MASC05927	0.12	0.16	0.12	0.12	0.11	0.16	0.13	0.2	0.13	0.14	0.13	0.15	0.12	0.11	0.13	0.12	0.12	0.11	0.13
RND_8	MN2_17844494	NA	NA	NA	NA	0.11	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RND_8	MN5_25963543	0.15	0.12	0.13	0.13	0.14	0.11	0.13	0.14	0.15	0.14	0.15	0.15	0.13	0.11	0.16	0.12	0.12	0.14	0.12
RND_8	PHYB_2850	0.12	0.16	0.13	0.12	0.12	0.14	0.13	0.16	0.14	0.14	0.16	0.14	0.11	0.1	0.13	0.13	0.14	0.12	0.14
RND_9	MASC02020	NA	NA	NA	NA	0.12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RND_9	MASC05927	0.12	0.14	0.12	0.12	0.11	0.15	0.14	0.19	0.13	0.14	0.13	0.14	0.13	0.12	0.14	0.12	0.12	0.11	0.13
RND_9	MN2_16080260	0.12	0.14	0.14	0.13	0.12	0.15	0.14	0.15	0.14	0.12	0.14	0.14	0.14	0.11	0.15	0.12	0.12	0.11	0.15
RND_9	MN4_17803780	0.11	0.13	0.13	0.14	0.16	0.13	0.12	0.13	0.16	0.14	0.13	0.13	0.15	0.14	0.14	0.12	0.12	0.13	0.13
RND_9	MN5_25963543	0.15	0.12	0.13	0.13	0.15	0.11	0.13	0.13	0.15	0.14	0.15	0.15	0.13	0.11	0.16	0.11	0.11	0.13	0.12
SOL_0	MN1_19506032	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6.83	NA	NA	NA	NA	NA	NA	NA
SOL_1	MN2_14406873	9.96	6.62	8.28	10.18	9.56	9.01	9.88	6.46	7.3	7.97	7.6	7.48	7.25	9.95	8.35	8.88	8.49	9.88	7.64
SOL_2	MN2_14406873	11.56	8.25	10.54	12.33	12.62	10.72	11.83	8.36	8.83	10.05	9.74	8.47	9.47	12.83	10.39	10.97	10.5	12.13	9.88
SOL_2	SAR1_183	11.4	8.45	10.38	12.22	12.24	10.58	11.85	8.57	9.1	10.1	10.01	9.13	9.62	12.83	10.66	10.91	10.49	11.43	10.27

Table B.2: Parental of Origin effects.
CPT = Compactness. ISO=Isotropy
PRM = Perimeter. RND = Roundness
PCx = Principal Component x ECC = Eccentricity

B.2 Genes within potential QTL

Marker	Gene	Description
MN1_1945105	AT1G06380	Ribosomal protein L1p/L10e family;
MN1_1945105	AT1G06370	pseudogene, similar to putative glycolipid alpha-mannosyltransferase
MN1_2548265	AT1G08130	DNA ligase 1. Key role in both DNA replication and excision repair pathways.
SGCSNP10165	NA	#N/A
MASC04211	AT1G43886	transposable element gene;
PERL0147872	AT1G47565	transposable element gene;
MN1_19506032	AT1G52360	Coatomer, beta' subunit;
MN1_592760	AT1G02720	Encodes a protein with putative galacturonosyltransferase activity.
MN1_1844838	AT1G06080	homologous to delta 9 acyl-lipid desaturases of cyanobacteria and acyl-CoA desaturases of yeast and mammals. expression down-regulated by cold temperature.
MN1_4041372	AT1G11960	ERD (early-responsive to dehydration stress) family protein;
FKF1_606	AT1G68050	Encodes FKF1. Regulates transition to flowering.
MASC00497	AT1G54410	dehydrin family protein;
MASC00290	AT1G56430	Encodes a protein with nicotianamine synthase activity.
MN1_21944762	AT1G59720	Pentatricopeptide Repeat Protein.
MN1_28584258	AT1G76160	SKU5 similar 5 (sks5);
NMSNP1_21401646	AT1G57770	FAD/NAD(P)-binding oxidoreductase family protein;
NMSNP2_16908351	AT2G40460	Major facilitator superfamily protein;
MASC05927	AT2G26240	Transmembrane proteins 14C;

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable

Marker	Gene	Description
MASC05920	AT2G26300	Encodes an alpha subunit of a heterotrimeric GTP-binding protein. Recessive mutant alleles have complex phenotypes including: reduced brassinolide response, reduced cell divisions, round leaves, short hypocotyls.
ER_472	AT2G26330	Homologous to receptor protein kinases. Involved in specification of organs originating from the shoot apical meristem.
MN2_10134400	AT2G23790	Protein of unknown function (DUF607);
MASC05841	AT2G33990	IQ-domain 9 (iqd9);
MASC06034	AT2G35585	unknown protein;
MASC02928	AT2G22910	N-acetyl-l-glutamate synthase 1 (NAGS1);
MN2_18034146	AT2G43410	FPA regulates flowering time in Arabidopsis independent of daylength. Mutations in FPA result in extremely delayed flowering.
MN2_14003409	NA	#N/A
HOS1_5954	AT2G39820	Translation initiation factor IF6;
PHYB_2850	AT2G18790	Red/far-red photoreceptor involved in the regulation of de-etiolation. Involved in the light-promotion of seed germination and in the shade avoidance response.
MASC05372	NA	#N/A
MASC02492	AT2G29930	F-box/RNI-like superfamily protein;
MN2_16684166	AT2G39950	unknown protein;
MN2_17792406	AT2G42720	FBD, F-box, Skp2-like and Leucine Rich Repeat domains containing protein;
MN2_17899626	AT2G43020	Encodes a polyamine oxidase.
MN2_17899626	AT2G43018	Upstream open reading frames (uORFs) can potentially mediate translational regulation of the largest, or major, ORF (mORF).

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable

Marker	Gene	Description
MN2_17304379	AT2G41480	Peroxidase superfamily protein;
MN2_17860020	AT2G42890	A member of mei2-like gene family.
ATC_2596	AT2G27570	P-loop containing nucleoside triphosphate hydrolases superfamily protein;
MASC02020	AT2G43290	Encodes calmodulin-like MSS3.
MN2_13605144	AT2G31960	encodes a protein similar to callose synthase
MN2_11300378	AT2G26550	Encodes a heme oxygenase-like protein lacking the conserved histidine residue at the active site that is usually involved in heme-iron coordination. It is unable to bind and degrade heme. Mutant analyses suggest a role in photomorphogenesis.
MN2_14406873	AT2G34100	unknown protein;
MN2_14406873	AT2G34110	forkhead-associated (FHA) domain-containing protein
MN2_14350717	AT2G33880	Encodes a protein with similarity to WUS type homeodomain protein. Required for meristem growth and development and acts through positive regulation of WUS.
MASC02512	AT2G32460	Member of the R2R3 factor gene family.
FES1_1177	AT2G33835	Encodes a zinc finger domain containing protein that is expressed in the shoot/root apex and vasculature, and acts with FRI to repress flowering.FES1 mutants in a Col(FRI+) background will flower early under inductive conditions.
MASC05384	AT2G36720	Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger domain;

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable

Marker	Gene	Description
RGA_1023	AT2G01570	Member of the VHIID/DELLA regulatory family. Putative transcriptional regulator repressing the gibberellin response and integration of phytohormone signalling. DELLAs repress cell proliferation and expansion that drives plant growth. Represses GA-induced vegetative growth and floral initiation. Rapidly degraded in response to GA. Involved in fruit and flower development.
MN2_7742622	AT2G17790	Encodes a protein with similarity to yeast VPS35 which encodes a component of the retromer involved in retrograde endosomal transport.
PHYB_5215	AT2G18790	Red/far-red photoreceptor involved in the regulation of de-etiolation. Involved in the light-promotion of seed germination and in the shade avoidance response.
MN2_16837474	AT2G40290	Encodes an eIF2alpha homolog that can be phosphorylated by GCN2 in vitro.
MASC06116	NA	#N/A
MASC02949	AT2G25680	Encodes a high-affinity molybdate transporter.
MN2_347772	AT2G01810	RING/FYVE/PHD zinc finger superfamily protein;
MN2_14801460	AT2G35100	Putative glycosyltransferase, similar to other CAZy Family 47 proteins.
MN2_12187411	AT2G28490	RmlC-like cupins superfamily protein;
MASC02995	AT2G20830	transferases;
FDP_733	AT2G17770	Encodes a paralog of bZIP transcription factor FD. This protein interacts with FD and FT.
MASC03019	AT2G23260	UDP-glucosyl transferase 84B1 (UGT84B1);
MN2_12435349	AT2G28940	Protein kinase superfamily protein;
MN2_15565512	AT2G37040	Encodes PAL1, a phenylalanine ammonia-lyase.

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable

Marker	Gene	Description
NMSNP2_14697188	AT2G34820	basic helix-loop-helix (bHLH) DNA-binding superfamily protein;
MASC05584	AT2G21350	RNA-binding CRS1 / YhbY (CRM) domain protein;
MN2_9256220	AT2G21620	Encodes gene that is induced in response to dessication;
MASC06022	AT2G40435	BEST Arabidopsis thaliana protein match is: transcription regulators (TAIR:AT3G56220.1);
MN2_14949589	AT2G35600	Belongs to five-member BRX gene family.
HOS1_1788	AT2G39810	A novel protein with a RING finger motif near the amino terminus. Negative regulator of cold responses.
MN2_9653239	AT2G22680	Zinc finger (C3HC4-type RING finger) family protein;
SAR1_183	AT2G33120	Encodes a member of Synaptobrevin-like protein family. Also known as VESICLE-ASSOCIATED MEMBRANE PROTEIN 722 (VAMP722). Required for cell plate formation.
MN2_13265590	AT2G31110	Encodes a member of the TBL (TRICHOME BIREFRINGENCE-LIKE) gene family containing a plant-specific DUF231 (domain of unknown function) domain.
MN2_13265590	AT2G31100	alpha/beta-Hydrolases superfamily protein;
MASC09221	NA	#N/A
PHYB_4171	AT2G18790	Red/far-red photoreceptor involved in the regulation of de-etiolation. Involved in the light-promotion of seed germination and in the shade avoidance response.
MN2_14407001	AT2G34110	forkhead-associated (FHA) domain-containing protein
MN2_14407001	AT2G34100	unknown protein;

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable

Marker	Gene	Description
MN2_17844494	AT2G42870	Encodes PHYTOCHROME RAPIDLY REGULATED1 (PAR1). Up regulated after simulated shade perception. Acts in the nucleus to control plant development and as a negative regulator of shade avoidance response. Transcriptional repressor of auxin-responsive genes SAUR15 (AT4G38850) and SAUR68 (AT1G29510).
MN2_16080260	AT2G38370	Plant protein of unknown function (DUF827);
NMSNP3_18379643	AT3G49550	unknown protein;
MN3_2236721	AT3G07060	embryo defective 1974 (emb1974);
NMSNP3_17381469	AT3G47170	HXXXD-type acyl-transferase family protein;
MASC04560	NA	#N/A
MASC02733	AT3G23580	Encodes one of the 3 ribonucleotide reductase (RNR) small subunit genes (RNR2A). Critical for cell cycle progression, DNA damage repair and plant development.
MN3_2967877	AT3G09670	Tudor/PWWP/MBT superfamily protein;
NMSNP3_2216922	AT3G07010	Pectin lyase-like superfamily protein;
NMSNP3_18980664	AT3G51070	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein;
NMSNP3_18980664	AT3G51075	Potential natural antisense gene, locus overlaps with AT3G51070
MN3_22146586	NA	#N/A
MN3_4141103	AT3G12970	unknown protein;
NMSNP3_21989468	NA	#N/A
MN3_15977654	AT3G44274	unknown pseudogene
NMSNP5_6416385	AT5G19130	GPI transamidase component family protein / Gaa1-like family protein;
MASC01180	AT5G61820	molecular function unknown;
NMSNP5_6523120	AT5G19360	member of Calcium Dependent Protein Kinase

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable

Marker	Gene	Description
MASC04437	AT5G63840	radial swelling mutant shown to be specifically impaired in cellulose production. Encodes the alpha-subunit of a glucosidase II enzyme.
MN5_25963543	AT5G64930	Regulator of expression of pathogenesis-related (PR) genes. Participates in signal transduction pathways involved in plant defense (systemic acquired resistance - SAR).
MASC01444	AT5G59320	Predicted to encode a PR (pathogenesis-related) protein. Belongs to the lipid transfer protein (PR-14) family
NMSNP5_13614633	AT5G35390	Encodes a member of the receptor-like kinase family of genes.
NMSNP5_14661352	AT5G37060	member of Putative Na+/H+ antiporter family
MN5_24996889	AT5G62180	carboxyesterase 20 (CXE20);
MASC04571	AT5G57530	xyloglucan endotransglucosylase/hydrolase 12 (XTH12);
MASC01545	AT5G57760	unknown protein;
NMSNP5_13272366	AT5G34965	transposable element gene;
MN4_11579839	AT4G21820	calmodulin binding
MN4_11580143	AT4G21820	calmodulin binding
NMSNP4_15765120	AT4G32680	unknown protein;
NMSNP4_15765120	AT4G32690	Encodes a hemoglobin (Hb) with a central domain similar to the 'truncated Hbs of bacteria, protozoa and fungi.
MASC04642	NA	#N/A
MN4_15087653	AT4G30990	ARM repeat superfamily protein;
MN4_17803780	AT4G37870	Encodes a phosphoenolpyruvate carboxykinase that localizes to the cytosol.
NMSNP4_17038400	AT4G36000	Pathogenesis-related thaumatin superfamily protein;

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable

Marker	Gene	Description
MN4_17442969	AT4G37000	Mutants have spontaneous spreading cell death lesions and constitutive activation of defenses in the absence of pathogen infection. Its product was shown to display red chlorophyll catabolite reductase (RCCR), which catalyzes one step in the breakdown of the porphyrin component of chlorophyll. The enzyme was further assessed to be a Type-1 (pFCC-1-producing) RCCR.Upon P. syringae infection, ACD2 localization shifts from being largely in chloroplasts to partitioning to chloroplasts, mitochondria, and to a small extent, cytosol. Overexpression of ACD2 delayed cell death and the replication of P. syringae.
PHYD_2290	AT4G16250	Encodes a phytochrome photoreceptor with a function similar to that of phyB that absorbs the red/far-red part of the light spectrum and is involved in light responses.
MN4_11488359	AT4G21605	pseudogene, hypothetical protein
MASC03336	AT4G32020	unknown protein;

Table B.3: Set of Markers, Genes and Gene Descriptors. Markers are the peak SNPs resulting from the QTL mapping using ER_472 as Covariable